

This correspondence is being deposited with the United States Postal Service as Express Mail addressed to: Box Patent Application, Assistant Commissioner for Patents, Washington, DC 20231, on October 12, 2001, Express Mail Receipt No. EL928548 091 US.

1 MULTI-SERVICE SEGMENTATION AND REASSEMBLY DEVICE
2 THAT MAINTAINS ONLY ONE
3 REASSEMBLY CONTEXT PER ACTIVE OUTPUT PORT
4

5 Bidyut Parruck
6 Chulanur Ramakrishnan
7

8 TECHNICAL FIELD

9 This invention relates to multi-service network communications, such as, for
10 example, line card circuitry disposed within routers and switches.
11

12 BACKGROUND INFORMATION

13 Figure 1 (Prior Art) is a diagram of a part of the Internet 1. The Internet, loosely
14 defined, is a collection of networks that are interconnected by devices called
15 “routers”. In the illustration, the Internet 1 involves seven networks N1-N7 and five
16 routers R1-R5. A protocol called the Internet Protocol (IP) is used to communicate a
17 message from a source device (a node) on one network to a destination device (a
18 node) on another network. The message is broken up into pieces and each of these
19 pieces is packaged into what is called an “IP packet”. These packets may be of
20 varying lengths. The IP packets of the message are then sent from the source to the
21 destination from one network to the next via the routers. The various IP packets can
22 take different paths to get from the source to the destination. When all the IP
23 packets arrive at the destination, they are reassembled to recreate the original
24 message.

25 This high level IP message can be transported across an individual network using
26 any one of many lower level protocols. Some of the protocols are packet-based
27 protocols, whereas others of the protocols are cell-based protocols. One packet-
28 based protocol used to transport IP is called Multi-Protocol Label Switching (MPLS).
29 In MPLS, each packet is encapsulated with an MPLS label by the first MPLS device

1 it encounters as it enters an MPLS network. The MPLS device is called an MPLS
2 edge router. The MPLS edge router analyses the contents of the IP header and
3 selects an appropriate MPLS label with which to encapsulate the packet. MPLS
4 packets therefore have varying lengths in the same way that IP packets do. At all
5 the nodes within the network subsequent to the edge router, the MPLS label (and
6 not the IP header) is used to make the forwarding decisions for the packet. Paths
7 through particular nodes in the network are setup from edge to edge, the label
8 defining the particular path its packet will take. Finally, as an MPLS labeled packet
9 leaves the network via an edge router, the edge router removes the MPLS label.

10 One cell-based lower level protocol used to transport IP over a network is the
11 Asynchronous Transfer Mode (ATM) protocol. In ATM, all packets are of equal
12 length. They are therefore called "cells". A large IP packet is transported over an
13 ATM network by segmenting the large IP packet into a plurality of smaller pieces.
14 Each of the smaller pieces is packaged to become an ATM cell. The ATM cells are
15 then transported across the ATM network. When the ATM cells reach the edge of
16 the ATM network, their payloads are reassembled to reform the large IP packet. In
17 Figure 1, networks N1, N5 and N3 are cell-based ATM networks. Networks N2, N6,
18 N4 and N7 are packet-based MPLS networks.

19 In the example of Figure 1, networks N3 and N4 are OC-192 high-speed
20 networks adapted to carry traffic over long distances. Router R2 at one end of
21 network N3 may, for example, be located in San Francisco whereas router R4 at the
22 other end of network N3 may be located in New York. Such high-speed long
23 distance networks are often called the "backbone" of the Internet.

24 In the example of Figure 1, individual users U1-U10 are coupled to the Internet
25 via local area networks. Networks N1, N2 and N7 are local area networks. In one
26 example where the network is a corporate network serving an office building, the
27 users are corporate employees in a building. In an example where the network is a
28 network operated by an Internet Service Provider (ISP), the users are individual
29 customers that pay the ISP to gain access to the Internet.

30 Consider the situation where users on networks N1 and N2 issue IP messages
31 that are destined to go to destinations to the right side of the backbone such that the
32 messages should go through one of the two back bone networks N3 and N4. In

1 such a case, the IP traffic from networks N1 and N2 is aggregated and supplied to
2 the router access point on the appropriate backbone network. A portion of the
3 Internet called the "Metropolitan Area" performs this function. In the illustration, the
4 metro area includes a router R1 used for aggregating traffic from networks N1 and
5 N2, and for routing that information to the appropriate one of backbone networks N3
6 and N4.

7 Figure 2 (Prior Art) is a more detailed view of router R1. Router R1 includes line
8 cards 2-3 for interfacing to ATM networks, other line cards 4 and 5 for interfacing to
9 MPLS networks, and a switch fabric 11. ATM line card 3 is coupled to ATM network
10 N5 such that router R1 can communicate with backbone network N3 via network N5.
11 Similarly, MPLS line card 5 is coupled to MPLS network N6 such that router R1 can
12 communicate with backbone network N4 via network N6. ATM line card 2 is coupled
13 to ATM network N1 via OC-12 fiber optic link 6, SONET multiplexer 7, higher speed
14 OC-48 fiber optic link 8, and SONET multiplexer 9. MPLS line card 4 is coupled to
15 MPLS network N2 via OC-12 fiber optic link 10, SONET multiplexer 7, higher speed
16 OC-48 fiber optic link 8, and SONET multiplexer 9. SONET multiplexer 7 performs
17 time division multiplexing (TDM) to modulate both ATM traffic from network N1 as
18 well as packet MPLS traffic from network N2 onto the same wavelength channel
19 transmitted over the same fiber optic link 8. SONET multiplexer 9 performs the
20 inverse function of time demultiplexing the signal on fiber optic link 8 to extract and
21 separate the ATM traffic from the MPLS traffic.

22 Router R1, when it receives an IP message from one of networks N1 or N2,
23 determines whether to forward the message on the message's "next hop" to router
24 R2 or R3. In this way IP network information from the users is aggregated in the
25 metro area and is directed to the correct backbone network for communication
26 across long distances to the destination.

27 A problem may exist if one of the local area networks coupled to router R1 is
28 disconnected or if the type of traffic on that network is changed from MPLS packet
29 traffic to ATM cell traffic or visa versa. Consider the situation in which ATM network
30 N1 ceases business operations. In that case, the operator of router R1 will likely
31 want to disconnect network N1 from its SONET multiplexer 7 and to couple in the
32 network of another paying customer. For example, the operator may want to

1 disconnect ATM network N1 and to connect in its place MPLS network N7. If this is
2 done, however, MPLS traffic would be received on ATM line card 2. ATM line card 2
3 is not suitable for coupling to an MPLS network. Consequently, ATM line card 2
4 may have to be disconnected and a suitable MPLS line card substituted in its place.
5 With the expansion of the Internet and with advances in IP switching technology, it
6 appears that the proportion of packet networks to ATM networks is increasing.
7 Accordingly, as more and more of the networks coupled to a router such as router
8 R1 migrate from one type of traffic to the other, more and more of the line cards of
9 the router will have to be replaced. This is undesirable. A solution is desired whereby a
10 smooth and easy migration from one type of traffic to the next is possible without the
11 removal of line cards or the physical manipulation of the router.

12 Figure 3 is a diagram of one possible approach to the problem involving a line
13 card 12 that handles both ATM and packet traffic. Line card 12 is coupled to a
14 switch fabric of a router by interface 13. Cell and packet traffic received from fiber
15 optic cable 14 and transmitted on fiber optic cable 15 are time division
16 multiplexed/demultiplexed by TDM device 16. Cell traffic is handled by integrated
17 circuit device 17. Packet traffic is handled by integrated circuit device 18. As the
18 relative amounts of cell traffic to packet traffic change, the same line card can be
19 used.

20

21 SUMMARY

22 Although the line card set forth in Figure 3 is satisfactory for some applications,
23 the general approach of Figure 3 involves substantial cost. The non-recurring
24 engineering costs associated with developing an integrated circuit can be quite high.
25 Even though the development of a particular integrated circuit may make technical
26 sense for a given application, it may be economically unreasonable to do so where
27 production volumes of the integrated circuit would be low. For each line card of
28 Figure 3, there is one integrated circuit for handling ATM traffic and one integrated
29 circuit for handling packet traffic. Developing a single integrated circuit having the
30 functionality of both the ATM device and the packet device might involve less non-
31 recurring engineering costs than developing two different integrated circuits, but the
32 integration would likely result in an undesirably large integrated circuit. Parts of such

1 a single integrated circuit may see little use in certain circumstances. Consider the
2 situation in which the mix of ATM traffic to packet traffic shifts to where there is little
3 or no ATM traffic. The single integrated circuit would involve a data path and
4 associated circuitry for handling ATM traffic that is underutilized or is not used at all.
5 Providing this extra unnecessary circuitry on the single integrated circuit would
6 constitute an unnecessary expense. It would therefore make the line card solution
7 afforded by the single integrated circuit unnecessarily expensive. It would be
8 preferable to get more use out of the circuitry provided on the integrated circuit in
9 order to reduce costs.

10 Not only might some circuitry be underutilized, but also so might other circuitry
11 become overburdened. In the above situation, for example, more and more
12 processing responsibilities would be required of the packet handling circuitry. If the
13 processing capability of the packet handling circuitry is sized to accommodate a
14 particular load, then increasing the load beyond that load by more and more of the
15 traffic shifting to packet traffic may result in the line card being inadequate to handle
16 the required traffic.

17 In one novel aspect, the same circuitry on a single Multi-Service Segmentation
18 And Reassembly (MS-SAR) integrated circuit handles both cell traffic and packet
19 traffic. Rather than there being a first data path through the integrated circuit for cell
20 processing, and another data path through the integrated circuit for packet
21 processing, functional blocks along a single data path process cell and/or packet
22 information that passes through the integrated circuit through the single data path.
23 Individual flows passing through the single data path are processed in accordance
24 with the traffic type of the individual flow. Any mix of cell to packet traffic can be
25 accommodated, thereby enabling a smooth migration from one traffic type to
26 another. The MS-SAR can handle dynamic changes in the relative amounts of cell
27 and packet traffic. Production costs associated with the integrated circuit are
28 reduced because the same functional blocks are used, albeit in different ways, to
29 process different types of flows.

30 In another novel aspect, production volumes may be increased in order to realize
31 economies of scale and to reduce per part cost. To allow production volumes to be
32 increased, the very same MS-SAR integrated circuit is usable not only both as an

1 ingress device and an egress device, but also with either a packet-based switch
2 fabric or a cell-based switch fabric. By providing a single MS-SAR integrated circuit
3 that functions as either an ingress device or an egress device as compared to a
4 device that can function as just one or the other, the number of applications for the
5 MS-SAR integrated circuit is increased. By providing a single MS-SAR integrated
6 circuit that can work with either a packet-based switch fabric or a cell-based switch
7 fabric, the number of applications for the single MS-SAR integrated circuit is
8 increased as compared to the same device that could only function with one type of
9 switch fabric.

#3 10 In another novel aspect, a single MS-SAR integrated circuit involves a lookup
11 block, a segmentation block, and a reassembly block, wherein traffic flows through a
12 single data path through the lookup block, and then through the segmentation block,
13 and then through the reassembly block. By using two identical such MS-SAR
14 devices, one in an ingress mode and another in an egress mode, information
15 received onto the ingress MS-SAR in either ATM or packet format can be
16 communicated through either a packet-based or a cell-based switch fabric, and can
17 be output from the egress MS-SAR in either ATM or packet format. Information
18 communicated in AAL5 adaptation cells received onto the ingress MS-SAR can be
19 reassembled and output in packet format on the egress MS-SAR. Packet format
20 information received onto the ingress MS-SAR can be segmented and output in the
21 form of AAL5 adaptation cells from the egress MS-SAR. The versatility of the single
22 MS-SAR to handle many different traffic types using the same circuitry further
23 increases the number of applications for which the integrated circuit can be used.

#4 24 In another novel aspect, individual flows are processed in different ways by an
25 egress MS-SAR. An indication of the type of egress processing to be done by an
26 MS-SAR on a flow is embedded into the flow that is received from a switch fabric
27 onto the egress MS-SAR. The egress MS-SAR reads the embedded indication and
28 performs the type of egress processing indicated. In one embodiment, the indication
29 of the type of egress processing is contained in a switch header (the switch header
30 can be either a switch header for a cell-based switch fabric or for a packet-based
31 switch fabric), the switch header being added by a first MS-SAR functioning in an
32 ingress mode, the first MS-SAR and the second MS-SAR being substantially

1 identical integrated circuits. In one embodiment, information on how to locate the
2 indication of the type in the information coming into the egress MS-SAR is provided
3 to the egress MS-SAR for each logical input port of the egress MS-SAR. The egress
4 MS-SAR uses: 1) the input port number of the flow, and 2) the information on how to
5 locate the indication for a given input port, to locate the indication in the information
6 coming in from the switch fabric.

7 In accordance with known ATM adaptation layer protocols, packet data can be
8 transmitted over an ATM network by segmenting the packet into smaller pieces and
9 then transmitting each smaller piece over the ATM network in the form of an ATM
10 cell. After transmission across the ATM network, the data payloads of the individual
11 ATM cells are recovered and are reassembled into the original packet. This
12 segmentation and reassembly process has been traditionally performed on a line
13 card by reassembling packets as the individual ATM cells are received onto the line
14 card. A reassembly context is maintained for each packet being reassembled. This
15 reassembly context may, for example, include a composite cyclic redundancy check
16 (CRC) value that is modified as each ATM cell carrying a portion of the packet is
17 received and processed. The CRC value calculated on the data portions received is
18 checked against a CRC transmitted in a trailer of the last ATM cell to verify that the
19 data carried by the numerous ATM cells has not been corrupted in the segmentation
20 and reassembly process. Once a packet has been checked and reassembled, is it
21 buffered into a payload memory on the line card. If, for example, the line card were
22 to support the simultaneous reassembly of one million packets, then the line card
23 would have to be able to store one million reassembly contexts. This would involve
24 a large amount of memory.

25 In another novel aspect, packets to be reassembled on the line card in such an
26 adaptation layer process are not reassembled before being buffered on the line card.
27 Rather, the payloads of the individual cells are buffered in payload memory as cells.
28 The line card does not maintain a reassembly context for each such packet being
29 buffered. When the buffered packet information is to be output from the line card,
30 the cell payloads corresponding to the packet are read out of payload memory and
31 the cell payloads are reassembled to form the packet. In this way, a large number of
32 packets being received onto the line card do not need to be simultaneously

1 reassembled, but rather the number of packets being simultaneously reassembled
2 can be set to have a smaller maximum. In one embodiment, one million flows to be
3 reassembled can be received at one time onto a line card, but only one packet per
4 active line card output port is reassembled at a time. By reducing the maximum
5 number of packets being simultaneously reassembled, the maximum number of
6 reassembly contexts to be stored on the line card is reduced. Reducing the number
7 of reassembly contexts to be stored reduces the amount of memory necessary and
8 thereby reduces line card costs.

9 Not only is a reassembly context involved the reassembly process, but a
10 segmentation context is also traditionally involved in the segmentation process.
11 Traditionally, packets to be segmented on a line card in accordance with an
12 adaptation layer process are received and stored into payload memory as packets.
13 When a packet is to be output, it is retrieved from payload memory and is
14 segmented into segments such that each of the segments forms the data payload of
15 an ATM cell. To be able to check the integrity of the data when the segments are
16 reassembled, a CRC is calculated on the packet at the time of segmentation and is
17 transmitted in a trailer that is included in the last ATM cell. For each such
18 segmentation process that is going on at the same time, a segmentation context
19 including a partial CRC value is maintained. If a large number such as a million
20 simultaneous output flows is to be supported by the line card, then the large number
21 of segmentation contexts must be stored on the line card. This involves a lot of
22 memory and consequently increases line card cost.

23 In another novel aspect, packets to be segmented in accordance with an
24 adaptation layer protocol are segmented on a per input port basis as they are
25 received onto the line card and prior to being buffered on the line card. The packets
26 are not buffered on the line card, but rather segments are buffered. Because only
27 one segmentation is performed at a time for a given line card input port, the
28 maximum number of simultaneous segmentations is limited to the number of input
29 ports. By limiting the maximum number of simultaneous segmentations to be
30 performed, the memory required to store the associated segmentation contexts is
31 reduced. In one embodiment, one million simultaneous flows can be processed, but

1 there are only sixty-four input ports. The amount of memory required for storing
2 segmentation contexts is therefore significantly reduced.

#7 3 In another novel aspect, an MS-SAR involves a data path such that data received
4 onto the MS-SAR passes through the data path and to a memory manager that
5 stores the data into payload memory. The data is then read out of the payload
6 memory and passes through the remainder of the data path to be output from the
7 MS-SAR. How and when the data is read out of payload memory is controlled by
8 control circuitry. The control circuitry controls the memory manager so that the
9 memory manager retrieves data from payload memory so that it will be output from
10 the MS-SAR in a manner controlled by the control circuitry. In one novel aspect, the
11 MS-SAR is partitioned into two integrated circuits such that the data path circuitry is
12 disposed on one integrated circuit and such that the control circuitry is disposed on
13 another integrated circuit. Partitioning the MS-SAR in this way facilitates future
14 increasing of data throughput rates without redesigning the MS-SAR. To increase
15 data throughput, for example from OC-192 rates to OC-768 rates, multiple data path
16 integrated circuits are disposed in parallel, each of the data path integrated circuits
17 being controlled by the same control integrated circuit. The control integrated circuit
18 has multiple control interfaces, one such interface for coupling to and controlling
19 each of the data path integrated circuits.

20 In another novel aspect, a router involves a first line card and a second line card.
21 Each of the first and second line cards involves an MS-SAR operating in the ingress
22 mode and an MS-SAR operating in the egress mode. The egress MS-SAR on the
23 second line card can become endangered of being overloaded if, for example, the
24 ingress MS-SAR on the first line card continues to send network information for a
25 flow to the egress MS-SAR on the second line card, but the egress MS-SAR on the
26 second line card is prevented from outputting that information, for example due to
27 congestion at the framer. Consequently, more and more of the network information
28 for the flow will end up having to be buffered by the egress MS-SAR of the second
29 line card. In one novel aspect, the ingress and egress MS-SAR devices of the first
30 line card are linked by a serial bus on the first line card, and the ingress and egress
31 MS-SAR devices of the second line card are linked by a serial bus on the second
32 line card. If the egress MS-SAR of the second line card is in danger of becoming

1 overloaded, then the egress MS-SAR of the second line card sends an indication of
2 this situation to the ingress MS-SAR of the second line card via the serial bus on the
3 second line card. The ingress MS-SAR of the second line card relays that
4 information to the first line card by outputting a special status switch cell. The
5 special status switch cell is transported across the switch fabric to the egress MS-
6 SAR of the first line card. The egress MS-SAR of the first line card detects the
7 special status switch cell, and relays the indication of the situation to the ingress MS-
8 SAR of the first line card via the serial bus on the first line card. In response to
9 receiving this indication from the serial bus, the ingress MS-SAR on the first line card
10 slows or stops outputting the information that is overburdening the egress MS-SAR
11 on the second line card.

12 In another novel aspect, an integrated circuit includes a reassembly circuit,
13 wherein in an ingress mode the reassembly circuit uses a flow ID to lookup a switch
14 header in an memory, and wherein in an egress mode the reassembly circuit uses a
15 flow ID to lookup a network protocol header (for example, ATM header or MPLS
16 header) in the memory.

17 These are but some of the novel aspects. Other novel structures and methods
18 are disclosed in the detailed description below. This summary does not purport to
19 define the invention. The invention is defined by the claims.

20 21 BRIEF DESCRIPTION OF THE DRAWINGS

22 Figure 1 (Prior Art) illustrates a part of the Internet where different traffic types are
23 aggregated.

24 Figure 2 (Prior Art) is a diagram of a router in the part of the Internet illustrated in
25 Figure 1.

26 Figure 3 is a diagram illustrative of an approach to solving a problem associated
27 with the network structure of Figure 1. Although the line card as illustrated may and
28 is intended to reflect a product designed by or being designed by Northern Telecom
29 Ltd., adequate details about this line card are not available to the inventors or the
30 assignee to state here in this patent document that what is shown in Figure 3 is prior
31 art.

1 Figure 4 is a diagram of a switching device (in this case a router) in accordance
2 with an embodiment of the present invention.

3 Figure 5 is a diagram of a line card in the router of Figure 4.

4 Figure 6 is a diagram that sets forth various application types that the router of
5 Figure 4 (involving a pair of MS-SAR devices, one operating in an ingress mode,
6 and the other operating in an egress mode) can carry out.

7 Figure 7 is a diagram illustrating ingress and egress application types involving a
8 cell-based switch fabric. These application types can be carried out by the router of
9 Figure 4.

10 Figure 8 is a diagram illustrating ingress and egress application types involving a
11 packet-based switch fabric. These application types can be carried out by the router
12 of Figure 4.

13 Figure 9 is a diagram of an example in accordance with an embodiment of the
14 present invention wherein a first flow is processed in accordance with ingress
15 application type 3 and egress application type 11, and wherein a second flow is
16 processed in accordance with ingress application type 0 and egress application type
17 8.

18 Figure 10 is a diagram illustrating an MS-SAR and associated memories in
19 accordance with an embodiment of the present invention.

20 Figure 11 is a simplified representation of an MPLS packet of flow #1 in the
21 example of Figure 9.

22 Figure 12 is a simplified diagram of chunks of flow #1 that are output by the
23 incoming SPI-4 interface block in the example of Figure 9.

24 Figure 13 is a diagram that illustrates information flow into an ingress MS-SAR in
25 the example of Figure 9.

26 Figure 14 is a representation of a port table in the lookup engine.

27 Figure 15A-15C are diagrams of the building of a per flow queue for flow #1 in the
28 example of Figure 9.

29 Figure 16 is a diagram of a dequeue memory location (stores the header pointer)
30 for one FID.

31 Figures 17 and 18 are diagrams of the first and second enqueue memory
32 locations (store the tail pointer) for one FID.

1 Figure 19 is a diagram of a memory location (stores a pointer to a buffer in the
2 body of a queue) for an intermediate BID in a linked list.

3 Figure 20 is a diagram of an ATM cell of flow #2 in the example of Figure 9.

4 Figure 21 is a diagram of a 56-byte chunk of flow #2 as output from the incoming
5 SPI-4 interface block of the ingress MS-SAR in the example of Figure 9.

6 Figure 22 is a diagram of a 64-byte chunk of flow #2 as output from the
7 segmentation block of the ingress MS-SAR in the example of Figure 9.

8 Figure 23 is a diagram of a per flow queue for flow #2 of the example of Figure 9.

9 Figure 24 is a diagram of the port calendar in the reassembly block in the MS-
10 SAR.

11 Figures 25 and 26 are diagrams of the port empty and port full registers in the
12 reassembly block in the MS-SAR.

13 Figure 27 is a diagram that illustrates information flow from payload memory out
14 of the ingress MS-SAR in the example of Figure 9.

15 Figure 28 is a diagram that illustrates the three switch cells of flow #1 in the
16 example of Figure 9.

17 Figure 29 is a diagram of the switch cell for flow #2.

18 Figure 30 is a diagram that illustrates the general flow of information into egress
19 MS-SAR 200 in the example of Figure 9.

20 Figure 31 is a diagram of the first switch cell for flow #1 in the example of Figure
21 9.

22 Figure 32 is a diagram that illustrates the general flow of information out of
23 egress MS-SAR 200 in the example of Figure 9.

24 Figure 33 is a diagram that illustrates the format of one FID entry in the header
25 table of the MS-SAR.

26 Figure 34 illustrates three 64-byte chunks of flow #1 as the chunks pass from
27 reassembly block 205 to outgoing SPI-4 interface block 206 in the example of Figure
28 9.

29 Figure 35 illustrates the MPLS packet of flow #1 as output from framer 142 in the
30 example of Figure 9.

31 Figure 36 illustrates the ATM cell of flow #2 as output from reassembly block 205
32 in the example of Figure 9.

1 Figure 37 illustrates the ATM cell of flow #2 as output form outgoing SPI-4
2 interface block 206 of egress MS-SAR 200 in the example of Figure 9.

3 Figure 38 illustrates an example of application types 5, 6, 14 and 13.

4 Figure 39 illustrates the processing of flow #1 in accordance with ingress
5 application type 5 in the example of Figure 38.

6 Figure 40 illustrates the processing of flow #2 in accordance with ingress
7 application type 6 in the example of Figure 38.

8 Figure 41 illustrates the processing of flow #1 in accordance with egress
9 application type 14 in the example of Figure 38.

10 Figure 42 illustrates the processing of flow #2 in accordance with egress
11 application type 13 in the example of Figure 38.

12 Figure 43 illustrates an example wherein a flow is processed on an ingress MS-
13 SAR in accordance with ingress application type 1 and is processed on an egress
14 MS-SAR in accordance with egress application type 9.

15 Figure 44 illustrates an example wherein a flow is processed on an ingress MS-
16 SAR in accordance with ingress application type 2 and is processed on an egress
17 MS-SAR in accordance with egress application type 10.

18 Figure 45 illustrates an example wherein a flow is processed on an ingress MS-
19 SAR in accordance with ingress application type 4 and is processed on an egress
20 MS-SAR in accordance with egress application type 14.

21 Figure 46 illustrates an example wherein a flow is processed on an ingress MS-
22 SAR in accordance with ingress application type 6 and is processed on an egress
23 MS-SAR in accordance with egress application type 12.

24 Figure 47 illustrates an example wherein a flow is processed on an ingress MS-
25 SAR in accordance with ingress application type 6 and is processed on an egress
26 MS-SAR in accordance with egress application type 14.

27 Figure 48 is a diagram of an embodiment wherein MS-SAR functionality is
28 partitioned into two integrated circuit chips (a data path integrated circuit and a
29 control integrated circuit) such that multiple data path integrated circuits chips can be
30 used with one control integrated circuit chip to increase data path throughput.

31 Figure 49 is a diagram of a packet as it is output from the distribution integrated
32 circuit of Figure 48.

1 Figures 50 and 51 are diagrams that illustrate the building of a packet queue in
2 connection with the operation of the embodiment of Figure 48.

3 Figure 52 is a diagram that illustrates a technique for accessing certain
4 information stored in external memory in a reduced amount of time in connection
5 with the embodiment of Figure 48.

6 Figure 53 is a diagram that illustrates a serial bus that couples an egress MS-
7 SAR of a line card to an ingress MS-SAR of the same line card. The egress MS-
8 SAR can use the serial bus to backpressure the sending ingress MS-SAR.

9 Figure 54 is a block diagram on one particular embodiment of incoming SPI-4
10 interface block 201 of Figure 10.

11 Figure 55 is a diagram of input control block 801 of Figure 54.

12 Figure 56 is a diagram of output control block 803 of Figure 54.

13 Figure 57 is a block diagram of one particular embodiment of segmentation block
14 203 of Figure 10.

15 Figure 58 is a block diagram of one particular embodiment of memory manager
16 block 204 of Figure 10. Figures 58A and 58B together form a more detailed version
17 of Figure 58.

18 Figure 59 is a block diagram of one particular embodiment of reassembly block
19 205 of Figure 10. Figures 59A-59D together form a more detailed version of Figure
20 59.

21 Figures 60A-60D are diagrams that illustrate reassembly types carried out by the
22 reassembly block of Figure 59. The function of the reassembly block in each of
23 these reassembly types can be described at the functional level in Verilog, and
24 hardware circuitry realized from the Verilog using hardware synthesis software.

25 Figure 61 is a diagram of one particular embodiment of outgoing SPI-4 interface
26 block 206 of Figure 10. Figures 61A and 61B together form a more detailed version
27 of Figure 61.

28 Figure 62 is a diagram of CPU interface block 211 of Figure 10.

29

30 DETAILED DESCRIPTION

31 Figure 4 is a simplified diagram of a router 100 in accordance with an
32 embodiment of the present invention. Router 100 includes a plurality of line cards

1 101-104, a switch fabric 105 and a central processing unit (CPU) 106. The line
2 cards 101-104 are coupled to switch fabric 105 by parallel buses 107-114. In the
3 present example, each of parallel buses 107-114 is a 16-bit SPI-4, Phase II, LVDS
4 parallel bus operating at 400 MHz at a double data rate (DDR). CPU 106 is coupled
5 to line cards 101-104 by another parallel bus 131. In the present example, parallel
6 bus 131 is a 32-bit PCI bus. In this example, each of the line cards can receive
7 network communications in multiple formats. For example, line card 101 is coupled
8 to a fiber optic cable 115 such that line card 101 can receive from cable 115 network
9 communications at OC-192 rates in packets, ATM cells, and/or AAL5 cells. AAL5
10 cells are considered a type of ATM cell.

11 Line card 101 is also coupled to a fiber optic cable 116 such that line card 101
12 can output onto cable 116 network communications at OC-192 rates in packets,
13 ATM cells, and/or AAL5 cells. The fiber optic cables 117 and 118 across which line
14 card 103 communicates are also labeled in the diagram. All the line cards 101-104
15 in this example have substantially identical circuitry.

16 Figure 5 is a more detailed diagram of representative line card 101. Line card
17 101 includes OC-192 optical transceiver modules 119 and 120, two serial-to-parallel
18 devices (SERDES) 121 and 122, a framer integrated circuit 123, a IP classification
19 engine 124, two multi-service segmentation and reassembly devices (MS-SAR
20 devices) 125 and 126, static random access memories (SRAMs) 127 and 128, and
21 dynamic random access memories (DRAMs) 129 and 130. MS-SAR devices 125
22 and 126 are identical integrated circuit devices, one of which (MS-SAR 125) is
23 configured to be in an "ingress mode", the other of which (MS-SAR 126) is
24 configured to be in an "egress mode". Each MS-SAR device includes a mode
25 register that is written to by CPU 106 via bus 131. When router 100 is configured,
26 CPU 106 writes to the mode register in each of the MS-SAR devices on each of the
27 line cards so as to configure the MS-SAR devices of the line cards appropriately.

28

29 ROUTER AND LINE CARD:

30 Fiber optic cable 115 of Figure 4 can carry information modulated onto one or
31 more of many different wavelengths (sometimes called "colors"). Each wavelength
32 can be thought of as constituting a different communication channel for the flow of

1 information. Accordingly, optics module 119 converts optical signals modulated onto
2 one of these wavelengths into analog electrical signals. Optics module 119 outputs
3 the analog electrical signals in serial fashion onto a high-speed analog bus 132.
4 Serdes 121 receives this serial information and outputs it in parallel form to framer
5 123 via high-speed parallel bus 133. Framer 123 receives the information from bus
6 133, frames it, and outputs it to classification engine 124 via another SPI-4 bus 134.
7 Classification engine 124 performs IP classification and outputs the information to
8 the ingress MS-SAR 125 via another SPI-4 bus 135. The ingress MS-SAR 125
9 processes the network information in various novel ways (explained below), and
10 outputs the network information to switch fabric 105 (see Fig. 4) via SPI-4 bus 107.
11 All the SPI-4 buses of Figures 4 and 5 are separate SPI-4, phase II, 400 MHz DDR
12 buses having sixteen bit wide data buses.

13 Switch fabric 105, once it receives the network information, supplies that
14 information to one of the line cards of router 100. Each of the line cards is identified
15 by a "virtual output port" number. Router 100 can include up to 256 line cards.
16 Accordingly, the virtual output port number has a range of from 0 to 255.

17 In the example of Figure 4, switch fabric 105 may supply network information
18 received on fiber optic cable 115 to any one of fiber optic cables 116, 118, 136 or
19 137. It is one of the primary functions of router 100 to determine, based on the
20 certain information such as the intended destination of the network information, how
21 to route the information. In the case where the network information is in the IP
22 packet format, the router makes its decision on where to route the network
23 information based on an intended IP destination address present in the IP header of
24 each packet. In the case where the network information is in the ATM cell format,
25 the router makes its decision on where to route the network information based on a
26 virtual path identifier and virtual channel identifier (VPI/VCI) information in the ATM
27 header of each ATM cell.

28 If, for example, the network information received from fiber optic cable 115 is to
29 be output from fiber optic cable 118, then switch fabric 105 would direct that network
30 information to the "virtual output port" of line card 103. If, on the other hand, the
31 network information received from fiber optic cable 110 is to be output from fiber
32 optic cable 137, then switch fabric 105 would direct that network information to the

1 "virtual output port" of line card 104. To facilitate the rapid forwarding of such
2 network information through the switch fabric 105, network information passed to the
3 switch fabric 105 for routing is provided with a "switch header". The "switch header"
4 may be in a format specific to the manufacturer of the switch fabric of the router.
5 The switch header identifies the "virtual output port" to which the associated network
6 information should be routed. Switch fabric 105 uses the virtual output port number
7 in the switch header to route the network information to the correct line card.

8 It is, however, generally the router 100 that determines to which of the multiple
9 line cards particular network information will be routed. Accordingly, the router's
10 CPU 106 provisions lookup information in (or accessible to) the ingress MS-SAR
11 125 so that the MS-SAR 125 will append an appropriate switch header onto the
12 network information before the network information is sent to the switch fabric 105
13 for routing. The ingress MS-SAR 125 uses header information in the network
14 information to lookup the particular switch header specified by CPU 106 for the
15 particular network information. Once the MS-SAR 125 has found the switch header
16 that it is to append to the network information, MS-SAR 125 appends this switch
17 header to the network information and sends the network information with the switch
18 header on to the switch fabric 105 via SPI-4 bus 107.

19 Switch fabric 105 receives the network information and forwards it to the line card
20 identified by the particular "virtual output port" in the switch header. Consider the
21 example in which the network information is forwarded to the virtual output port of
22 line card 103 so that the network information from fiber optic cable 115 will be output
23 onto fiber optic cable 117. Because the structures of the various line cards are
24 identical in this example, the flow of network information through line card 103 is
25 explained as passing through the egress MS-SAR 126 of Figure 5. The network
26 information and switch header is received onto the egress MS-SAR 126 of the line
27 card that is identified by the virtual output port number in the switch header. The
28 egress MS-SAR 126 receives the network information, removes the switch header,
29 performs other novel processing (explained below) on the network information, and
30 outputs the network information to framer 123 via SPI-4 bus 138. Framer 123
31 outputs the network information to serdes 122 via high-speed parallel bus 139.
32 Serdes 122 converts the network information into serial analog form and outputs it to

1 output optics module 120 via high-speed analog bus 140. Output optics module 120
2 converts the information into optical signals modulated onto one wavelength
3 channel. This optical information is transmitted through fiber optic cable 116 (which
4 corresponds to cable 118).

5 Although each line card in the example of Figure 4 outputs to only one fiber optic
6 cable (i.e., has only one physical output port), this need not be the case. If, for
7 example, a high speed OC-192 optical transceiver and cable is coupled to a line
8 card, then the line card may only have one physical output port. If, on the other
9 hand, two lower speed optical transceivers (for example, OC-48 transceivers) and
10 cables are coupled to the line card, then the line card may have two physical output
11 ports (one being the first OC-48 transceiver and cable, the other being for the
12 second OC-48 transceiver and cable). The egress MS-SAR 126 outputs onto SPI-4
13 bus 138 network information to be transmitted along with an output port number from
14 0 to 63. (This "output port number" is not to be confused with the "virtual output port
15 number" used to identify line cards.) Groups of these 64 output port numbers are
16 mapped by framer 123 to the physical output ports provided on the line card. For
17 example, output ports 0-31 on SPI-4 bus 138 may be mapped to a first physical
18 output port (a first OC-48 output transceiver), whereas output ports 32-63 on SPI-4
19 bus 138 may be mapped to a second physical output port (a second OC-48 output
20 transceiver). In accordance with the SPI-4 bus protocol, there are 64 logical output
21 ports that egress MS-SAR 126 can specify on SPI-4 bus 138.

22 In a similar way, a line card may receive network information from more than one
23 optical input transceiver module and more than one fiber optic cable. Information
24 coming into a first optical input transceiver module (i.e., a first physical port) would
25 be supplied onto SPI-4 bus 134 by framer 123 with a logical input port number that
26 maps to the first optical input transceiver module onto which the information was
27 received. Similarly, information coming into a second optical input transceiver
28 module (i.e., a second physical port) would be supplied onto SPI-4 bus 134 by
29 framer 123 with a logical input port number that maps to the second optical input
30 transceiver module onto which the information was received. In accordance with the
31 SPI-4 bus protocol, there are 64 logical input ports that framer 123 can specify on

SPI-4 bus 134. The logical input port information on SPI-4 bus 134 flows through classification engine 124 to appear on SPI-4 bus 135.

APPLICATION TYPES:

A flow is a sequence of one or more packets or cells transmitted between a selected source and a selected destination, the flow generally representing a single session using a protocol. Each packet or cell in a flow is expected to have the same routing information according to the dictates of the protocol. The specific embodiment of the MS-SAR device described here can process up to one million flows simultaneously. Each of these one million flows is processed in accordance with any one of several "application types". In each application type, the ingress MS-SAR processes the flow in accordance with one of several "ingress application types" and the egress MS-SAR processes the flow in accordance with one of several "egress application types". Figure 6 sets forth, for each application type, the related ingress application type and the related egress application type. Figure 7 sets forth the ingress and egress application types when the switch fabric is a cell-based switch fabric. Figure 8 sets forth the ingress and egress application types when the switch fabric is a packet-based switch fabric.

EXAMPLE OF APPLICATION TYPES 0, 3, 8 AND 11:

Figures 9-37 illustrate an example of how two flows of different traffic types (flow #1 is an MPLS packet, flow #2 is an ATM cell) are received onto a first line card (in this example, line card 101), simultaneously pass through the ingress MS-SAR 125 on the first line card, pass through switch fabric 105 (in this case switch fabric 105 is a cell-based switch fabric), pass onto a second line card (in this example, line card 103), simultaneously pass through an egress MS-SAR 200 on the second line card 103, and are output by the second line card 103. In this example, flow #1 and flow #2 are both communicated to first line card 101 on the same wavelength channel transmitted on the same fiber optic cable 115. Similarly, flow #1 and flow #2 are both communicated from second line card 103 on the same wavelength channel transmitted on the same fiber optic cable 118.

1 In Figure 9, flow #1 is an MPLS packet flow passing into an ingress line card,
2 where the ingress line card supplies switch cells to the switch fabric. As indicated by
3 the lower left example of Figure 7, this type of flow is identified as ingress application
4 type 3. Ingress MS-SAR 125 therefore performs processing on flow #1 in
5 accordance with ingress application type 3. In Figure 9, flow #2 is an ATM cell flow
6 passing into an ingress line card, where the ingress line card supplies switch cells to
7 the switch fabric. As indicated by the upper left example of Figure 7, this type of flow
8 is identified as ingress application type 0. Ingress MS-SAR 125 therefore performs
9 processing on flow #2 in accordance with ingress application type 0.

10 Figure 10 is a more detailed diagram of the MS-SAR devices 125 and 200
11 present on line cards 101 and 103. Each MS-SAR device includes an incoming SPI-
12 4 interface block 201, a lookup engine block 202, a segmentation block 203, a
13 memory manager block 204, a reassembly and header-adding block 205, an
14 outgoing SPI-4 interface block 206, a per flow queue block 207, a data base block
15 208, a traffic shaper block 209, an output scheduler block 210, and a CPU interface
16 block 211. In one embodiment, the blocks within dashed line 212 are realized on a
17 first integrated circuit and the blocks within dashed line 213 are realized on a second
18 integrated circuit. In the event the functionality of the MS-SAR is broken into two
19 integrated circuits, a second CPU interface block 214 is provided so that CPU 106
20 can communicate with both integrated circuits. The MS-SAR interfaces to and uses
21 numerous other external memory integrated circuit devices 215-228 that are
22 disposed on the line card along with the MS-SAR.

23 In the example of Figure 9, the MPLS packet of flow #1 is received onto input
24 optics module 119 from fiber optic cable 115. The MPLS packet passes through
25 (see Figures 5 and 9) the input optics module 119, through serdes 121, and to multi-
26 service SONET framer 123. Framer 123 is a framer that includes a demapper for
27 separating ATM cells and MPLS packets that are received on the same wavelength
28 via input optics module 119. Framer 123 in one embodiment is a Ganges S19202
29 STS-192 POS/ATM SONET/SDH Mapper available from Applied Micro Circuits
30 Corporation, 200 Brickstone Square, Andover, MA 01810. Framer 123 outputs the
31 MPLS packet information of flow #1, along with a logical input port number indicative
32 of the physical input module from which the information was received, in 16-bit

1 pieces, onto SPI-4 bus 134. Classification engine 124 performs IP (Internet
2 Protocol) classification in the case where router 100 performs IP lookup.
3 Classification engine 124 may, in one embodiment, be a classification engine
4 available from Fast-Chip Incorporated, 950 Kifer Road, Sunnyvale, CA 94086. In
5 the particular example of Figure 9, the packet traffic (flow #1) comes into the router
6 in MPLS form and exits the router in MPLS form. Consequently, IP classification is
7 not performed in this embodiment.

8 Figure 11 is a simplified diagram of the MPLS packet 300 of flow #1 as it is
9 received onto ingress MS-SAR 125. MPLS packet 300 contains an MPLS header
10 301, and a data payload 303. An SPI-4 start delimiter 302 and an SPI-4 end
11 delimiter 304 frame the packet. MPLS header 301 includes the 20-bit MPLS label.
12 The MPLS header 301 is disposed in the packet between the layer two (L2) header
13 and the layer three (L3) IP header. The L2 header was removed by framer 123 such
14 that beginning of the packet as received onto ingress MS-SAR 125 is the MPLS
15 header.

16 Ingress MS-SAR 125 (see Fig. 10) receives the packet information in multiple
17 sixteen-byte bursts, sixteen bits at a time, via sixteen input terminals 198. When
18 incoming SPI-4 interface block 201 accumulates 64 bytes (block 201 accumulates
19 64 bytes in ingress mode and up to 80 bytes in egress mode), it sends the 64-byte
20 chunk to lookup block 202 via 64-bit wide data bus 317. In this way incoming SPI-4
21 interface block 201 breaks the packet up into 64-byte chunks and communicates the
22 chunks to lookup block 202.

23 Figure 12 illustrates the three 64-byte chunks into which this particular MPLS
24 packet 300 is broken as output from incoming SPI-4 interface block 201 to lookup
25 block 202. The incoming SPI-4 interface block 201 uses the start delimiter 302 and
26 the end delimiter 304 to identify the beginning and ending of the packet 300. The
27 incoming SPI-4 interface block 201 outputs two additional signals (an end-of-packet
28 signal and a start-of-packet signal) along with each of the 64-byte chunks. Figure 12
29 shows the end-of-packet (EOP) and start-of-packet (SOP) bits for each of the three
30 64-byte chunks.

31 Figure 13 shows information flow through ingress MS-SAR 125. CPU 106 has
32 previously placed lookup information into ingress MS-SAR 125 so that the

1 information in each MPLS packet header can be used by lookup block 202 to find: 1)
2 a particular flow ID (FID) for the flow that was specified by CPU 106, and 2) the
3 ingress application type. The ingress application type, once determined, is used by
4 other blocks of the ingress MS-SAR 125 to configure themselves in the appropriate
5 fashion to process the network information appropriately.

6 Only one traffic type (for example, MPLS packet or ATM cell) is permitted on
7 each logical input port of SPI-4 bus 135. The traffic type is assigned by CPU 106,
8 but there can only be one traffic type assigned at a given time for a given logical
9 input port. A "port table" in lookup block 202 contains, for each of the sixty-four
10 logical input ports, the allowed traffic type as specified by CPU 106. This allowed
11 traffic type information is provisioned in the port table before the flow starts by CPU
12 106.

13 Figure 14 is a conceptual diagram that shows the relationship of the information
14 in the "port table" in lookup block 202. For each logical input port there is a traffic
15 type defined. Lookup block 202 uses the incoming logical port number supplied by
16 framer 123 to lookup the traffic type allowed on the logical input port that packet 300
17 was received on. Each traffic type has a known format. Accordingly, once the traffic
18 type is known, the lookup block can locate in the header of the incoming flow the
19 particular information that will be used to lookup an associated FID found in a "FID
20 hash table". This "FID hash table" is stored in external memories 215 and 216 (see
21 Fig. 10).

22 In the present example of Figure 9, packet 300 is an MPLS packet. Lookup block
23 202 uses the traffic type found in the "port table" to locate the 20-bit MPLS label
24 within the first chunk of the packet. The MPLS label contains information similar to
25 source and destination addressing information. A hash generator in the lookup
26 block 202 uses the 20-bit MPLS label as a "hash key" to generate a "hash index".
27 This "hash index" is used as an address for the "FID hash table". The content of the
28 FID hash table location addressed contains both the FID and the ingress application
29 type.

30 The FID and ingress application type, once determined, are passed via 64-bit
31 wide data bus 318 (see Figure 10), one at a time, to segmentation block 203 along
32 with the consecutive 64-byte chunks of data. Segmentation block 203 stores these

1 64-bytes chunks on a per port basis. Each 64-byte chunk is stored in a buffer in
2 SRAM within segmentation block 203, and a pointer to the buffer is pushed onto a
3 FIFO of pointers for the input port. The FID and traffic type is also pushed onto the
4 FIFO with the pointer. In the present example, there are three 64-byte chunks in
5 flow #1. Segmentation block 203 calculates a cyclic redundancy check (CRC) value
6 for the data portion of the entire packet, and adds a trailer including this CRC such
7 that the trailer is at the end of the last 64-byte chunk. Segmentation block 203 adds
8 any pad that might be necessary between the end of the data of the last 64-byte
9 chunk and the trailer. Memory manager block 204 pops the FIFO of pointers such
10 that segmentation block 203 forwards the 64-byte chunks one at a time to memory
11 manager block 204 via a 128-bit wide bus 319.

12 Payload memory 217 contains a large number of 64-byte buffers, each buffer
13 being addressed by a buffer identifier (BID). Some of the 64-byte buffers of payload
14 memory 217 may be used and storing data, whereas others may not be storing data
15 (i.e., are "free"). Per flow queue block 207 maintains a linked list of the BIDs of
16 available buffers, called the "free buffer queue". The "free buffer queue" is stored in
17 external SSRAM 226 (see Figure 10). When memory manager block 204 receives
18 the first 64-byte chunk 305 of data associated with packet 300, memory manager
19 block 204 issues an "enqueue" command via enqueue command line 320 to per flow
20 queue block 207. This constitutes a request for the BID of a free buffer. Per flow
21 queue block 207 pops the free buffer queue to obtain the BID of a free buffer, and
22 forwards that BID to memory manager block 204 via lines 321. Memory manager
23 block 204 then stores the 64-byte chunk 305 (see Fig. 12) of data for packet 300 in
24 the buffer in payload memory 217 identified by the BID. The writing of the 64-byte
25 chunk of data is indicated in Figure 13 by the upward pointing heavy arrow 322 that
26 extends toward payload memory block 217.

27 Per flow queue block 207 also maintains a linked list (i.e., a "queue") of the BIDs
28 for the various 64-byte chunks of each flow that are stored in payload memory 217.
29 Such a linked list is called a "per flow queue". Figure 15A illustrates how per flow
30 queue block 207 builds the linked list for flow #1. Each linked list has a head pointer
31 and a tail pointer. A list of the head pointers for the one million possible FIDs is

"T02T01" E292650

1 stored in external SRAM 228 (see Figure 10). A list of the tail pointers for the one
2 million possible FIDs is stored in external SRAM 227.

3 The head pointer for the FID of flow #1 is set to point to the memory location that
4 stores the BID where the first chunk 305 is stored in payload memory 217. Because
5 there is only one chunk (C1) in the linked list for flow #1, the tail pointer is set to
6 point to the same location. This first chunk 305 of flow #1 is the start of packet 300
7 as identified by an SOP bit.

8 Next, the second 64-byte chunk 306 is received by the memory manager block
9 204. The memory manager block 204 again issues an enqueue command to per
10 flow queue block 207, again obtains a BID of a free buffer, and then writes the
11 second chunk 306 into the payload memory at the location identified by the BID.
12 The per flow queue block 207 pops the free buffer queue to remove the now-used
13 BID from the free buffer queue, and adds the BID for the second chunk 306 to the
14 linked list for the FID of flow #1. As illustrated in Figure 15B, the BID of the second
15 chunk 306 is added to the linked list for the FID of flow #1.

16 Next, the third 64-byte chunk 307 is received by the memory manager block 204.
17 The same sequence of storing the chunks into payload memory and adding the BID
18 to the linked list of flow #1 is carried out. Figure 15C illustrates the per flow queue
19 for flow #1 including pointers to these three chunks. In the case of chunk 307, this
20 chunk is the last chunk of the packet as indicated by the EOP bit. The linked list for
21 flow#1 is therefore complete.

22 Figure 16 is a diagram of a memory location in FID dequeue memory 228 (see
23 Figure 10) that stores the head pointer for a flow. Per flow queue block 207 stores in
24 the location, along with the BID head, other information it receives from memory
25 manager block 204 relating to the chunk. This information includes an EOP bit that
26 indicates whether the chunk is the last chunk of a packet, an SOP bit that indicates
27 whether the chunk is the first chunk of a packet, and the ingress application type.
28 Per flow queue block 207 also stores with the head pointer EFCI, CLP and OAM
29 bits. These bits were extracted by lookup block 202 from the header. There is one
30 memory location such as the one illustrated in Figure 16 for each per flow queue. In
31 the example of Figure 15C, chunk C1 is pointed to by a head pointer in one such
32 memory location.

1 Figures 17 and 18 are diagrams of two memory locations in FID enqueue
2 memory 227 (see Figure 10) that store a tail pointer. Per flow queue block 207
3 stores in these locations, along with a pointer to the BID tail, other information on the
4 chunk including the output port number from which the data will eventually be output,
5 the size of the per flow queue, and an indication of a time to live (a TTL bit). There
6 is one such pair of memory locations for each per flow queue. In the example of
7 Figure 15C, chunk C3 is pointed to by information in a pair of two such memory
8 locations.

9 Figure 19 is a diagram of a memory location in SRAM 226 that stores the pointer
10 for a chunk in a queue between the head and tail. Per flow queue block 207 stores
11 in this location a pointer to the next BID in the per flow queue. The location also
12 stores an EOP indication and an SOP indication. In the example of Figure 15C,
13 there is one such memory location for chunk C2.

14 In the example of Figure 9, a second flow (flow #2) of a different traffic type (ATM
15 cell) is received onto the same line card 101 that flow #1 was. In this example, flow
16 #2 is also received via the same fiber optic cable 115 and is modulated onto the
17 same wavelength channel that flow #1 was. The 53-byte ATM cell of flow #2 passes
18 through the same optics module 119, serdes 121 and framer 123. Framer 123
19 removes the fifth byte (the "HEC" Header Error Control byte) of the ATM header and
20 places the remaining 52-byte ATM cell onto the 16-bit SPI-4 bus 134. This
21 information passes through classification engine 124 to SPI-4 bus 135 such that the
22 52-byte ATM cell is received onto the incoming SPI-4 interface block 201 of ingress
23 MS-SAR 125 via sixteen SPI-4 input terminals 198 (see Fig. 10). Incoming SPI-4
24 interface block 201 receives the 52-byte ATM cell in multiples of 16-byte bursts, 16
25 bits at a time.

26 Figure 20 is a simplified diagram of the ATM cell 308 as received onto ingress
27 MS-SAR 125. ATM cell 308 is 52-bytes long. ATM cell 308 includes an ATM
28 header 309 and a data payload 311. An SPI-4 bus start delimiter 310 and end
29 delimiter 312 frame the ATM cell. Incoming SPI-4 interface block 201 receives the
30 ATM cell and supplies it to lookup block 202, sixty-four bits at a time, via 64-bit wide
31 data bus 317.

1 The internal data path within the MS-SAR from incoming SPI-4 interface block
2 201, through lookup block 202, and to segmentation block 203 is 64-bits (8-bytes)
3 wide. Incoming SPI-4 interface block 201 therefore adds a 4-byte pad 314 (see
4 Figure 21) to pad the 52-byte ATM cell up to 56-bytes (a multiple of 8 bytes) before
5 sending the ATM cell to lookup block 202.

6 Figure 21 is a diagram of ATM cell 308 as it is output from incoming SPI-4
7 interface block 201 to lookup block 202. Because the entire ATM cell 308 is
8 contained in one 64-byte chunk, the SOP and EOP signals output by the incoming
9 SPI-4 interface block 201 indicate both the start of packet (in this case the packet is
10 a cell) and end of packet (in this case the packet is a cell).

11 Lookup block 202 receives the 56-byte chunk for ATM cell 308, and from the
12 logical input port number looks up the traffic type from the "port table" (see Figure
13 14). The "port table" indicates that the traffic type is ATM cells. Lookup block 202
14 uses the traffic type to locate the 12-bit VPI and 16-bit VCI fields in the ATM cell
15 header. The hash generator of lookup block 202 uses the located VPI and VCI
16 information as a "hash key" to create a "hash index". This "hash index" is then used
17 to lookup the FID and ingress application type in the "FID table" stored in external
18 memories 215 and 216.

19 Once the FID and the ingress application type for flow #2 are determined, these
20 values are passed from lookup block 202 to segmentation block 203 via 64-bit bus
21 318 along with the 56-byte chunk. Segmentation block 203 adds an additional 8-
22 byte pad to pad the ATM cell up to 64-bytes, stores the 64-byte chunk into an SRAM
23 (not shown) in segmentation block 203, and pushes a pointer to that chunk onto its
24 FIFO of pointers for that input port. Memory manager block 204 pops the FIFO of
25 pointers such that segmentation block 203 supplies the 64-byte chunk from its
26 SRAM to memory manager block 203.

27 Figure 22 is a simplified diagram of the 64-byte chunk as output from
28 segmentation block 203 onto 128-bit bus 319. This 64-byte cell contains the 4-byte
29 pad 314 added by the incoming SPI-4 interface block 201 and the additional 8-byte
30 pad 315 added by segmentation block 203. When memory manager block 204
31 receives the 64-byte chunk, memory manager block 204 issues an enqueue
32 command via command line 320 to per flow queue block 207, obtains in return a

1 BID, and then stores the 64-byte chunk in payload memory 217 at the location
2 identified by the BID. Per flow queue block 207 pops the free buffer queue, thereby
3 removing the now-used BID from the free buffer queue, and adds the BID to a linked
4 list for flow #2.

5 Figure 23 is a diagram that illustrates the two linked lists. The linked list for flow
6 #1 (FID1) has BIDs for three chunks C1, C2 and C3, whereas the flow for flow #2
7 (FID2) has BIDs for one chunk C1. The chunk for the ATM cell is indicated as being
8 both the "start of packet" as well as the "end of packet".

9 Once the linked lists (queues) for flow #1 and flow #2 are formed, the linked lists
10 are popped (i.e., dequeued) in a particular way and order such that their associated
11 chunks in payload memory 217 are output from ingress MS-SAR 125 in a desired
12 fashion, taking into account user-programmable policing/metering parameters.
13 User-programmable parameters may include, for example, burst size, peak cell rate,
14 and sustainable cell rate.

15 The dequeue process starts with data base block 208 determining an output port
16 to service using a port calendar. This port calendar is located within data base block
17 208. Once the output port to service is selected, data base block 208 supplies the
18 output port to traffic shaper block 209 and to output scheduler block 210. The traffic
19 shaper block 209 and the output scheduler block 210 supply flow IDs back to data
20 base block 208. Data base block 208 selects one of the two flow IDs to dequeue for
21 that port. Data base block 208 gives priority to traffic shaper output over output
22 scheduler output such that only if the shaper has no FID to output for a given port
23 will the output scheduler be allowed to schedule an FID for that port. Either traffic
24 shaping is performed on an individual flow by traffic shaper block 209, or output
25 scheduling is performed on the flow by output scheduler block 210.

26 Traffic shaper block 209 performs traffic shaping on a per flow basis for up to one
27 million flows. On a particular flow, either a single leaky bucket shaping scheme is
28 used, a dual leaky bucket shaping scheme is used, or no shaping at all is used. The
29 shaper selects a peak rate or a sustained rate per flow ID depending on an
30 accumulated credit. Up to 1024 different programmable rates can be shaped
31 simultaneously. In one mode, the rate of incoming traffic with a particular flow ID is
32 measured (i.e., metered) and the flow ID of a violator is marked if the measured rate

1 is above a specific threshold programmed by CPU 106 for the flow ID. Up to 1024
2 different programmable thresholds can be metered simultaneously.

3 Output scheduler block 210 uses a weighted round-robin scheme to select a
4 quality of service of the selected port. Once the quality of service is selected, a flow
5 ID is selected based on a round-robin scheme.

6 When data base block 208 receives a flow ID from either traffic shaper block 209
7 or output scheduler block 210, data base block 208 generates a request to per flow
8 queue block 207 to issue a dequeue command to dequeue the flow ID. Per flow
9 queue block 207 accesses the per flow queue of the flow ID, determines the next
10 BID to dequeue, and outputs the BID in the form of a dequeue command to memory
11 manager block 204.

Per flow queue block 207 is programmable to cause the linked list of one flow to be output multiple times (i.e., multicast), each time with a different flow ID. Multicasting is performed by replicating the linked list of a flow, each replica linked list having its own flow ID. Per flow queue block 207 is also programmable to cause multiple flows to be combined into one flow (i.e., tunneling). Tunneling is performed by linking the tail of one linked list to the head of another so as to create one large composite linked list, the composite linked list having one flow ID.

19 In the present example, the linked list of flow #1 (see Figure 22) is dequeued first.
20 Per flow queue block 207 pops the BID of the first chunk C1 off the FID1 linked list
21 for flow #1 and forwards that BID to memory manager block 204 in a dequeue
22 command. The dequeue command contains the BID of the 64-byte chunk to be
23 read out of payload memory 217, as well as the ingress application type, an EOP bit,
24 an SOP bit, and the output port number (one of 64 logical output ports on SPI-4 bus
25 107). The dequeue command is sent via dequeue command line 323. The BID is
26 sent via BID lines 321. Per flow queue block 207 adds the now available BID to the
27 free buffer queue in external memory 226.

28 In response to receiving the dequeue command, memory manager block 204
29 retrieves the first chunk C1 identified by the BID and outputs that first chunk C1 to
30 reassembly block 205 via 128-bit data bus 324. Memory manager block 205
31 supplies to reassembly block 205 other control information including the FID of
32 chunk C1, the SOP and EOP bits, the ingress application type being performed on

1 flow #1, and a logical output port ID (PID) identifying the one of the 64 logical output
2 ports on SPI-4 bus 107 to which the chunk will be sent.

3 Reassembly block 205 uses the ingress application type to determine what type
4 of action it should perform on the associated chunk. In the present example, flow #1
5 is processed in accordance with ingress application type 3. External memory 218
6 (see Fig. 10) contains a "header table" 327. For each FID, the CPU 106 has stored
7 beforehand in header table 327 a switch header that reassembly block 205 is to
8 append to the data of the chunk before sending the data on to the switch fabric.
9 Accordingly, reassembly block 205 uses the FID to lookup in "header table" 327 the
10 "switch header" placed there for this FID by CPU 106. The switch header can be
11 either 8-bytes or 16-bytes depending on the requirements of the switch fabric. As
12 explained above, the "switch header" contains the "virtual port number" of the
13 particular egress line card that the switch fabric 105 wants the switch cell routed to.
14 In the present example of Figure 9, the egress line card is line card 103.
15 Accordingly, CPU 106 has placed the "virtual port number" of line card 103 into the
16 "switch header" associated with flow #1.

17 Reassembly block 205 also provides a special 4-byte "Azanda header". The
18 Azanda header is embedded in the switch cell as the last four bytes of the switch
19 header. This Azanda header contains information on how the switch cell should be
20 processed by the particular egress MS-SAR that receives the switch cell (in this
21 case MS-SAR 200). The Azanda header includes the egress application type to be
22 used to process the switch cell in the egress MS-SAR. The Azanda header also
23 includes the FID of the cell, an SOP bit, an EOP bit, and quality of service
24 information.

25 When reassembly block 205 receives the 64-byte chunk from memory manager
26 block 204, it stores the 64-byte chunk into a dual-port DATA_SRAM and stores the
27 "switch header" (including the "Azanda switch header") into a HDR_SRAM. The
28 DATA_SRAM (not shown) and the HDR_SRAM (now shown) are part of the
29 reassembly block 205. A pointer that points to the data in DATA_SRAM and to the
30 header in HDR_SRAM is pushed onto a queue for the particular logical output port.
31 In the present example, where the MS-SAR is operating in the ingress mode, there
32 is only one logical output port (i.e., the output of line card 101), consequently there is

1 only one output port queue. The pointer is therefore placed on this queue. The
2 output port queue is maintained in a Q_FIFO (not shown) that is also part of the
3 reassembly block 205.

4 Figure 24 is a conceptual diagram of a "port calendar" located in reassembly
5 block 205. Figure 25 is a diagram of a "port empty" register located in reassembly
6 block 205. Figure 26 is a diagram of a "port full" register located in reassembly block
7 205. Reassembly block 205 uses the port calendar and registers of Figures 24-26 to
8 determine how to pop the output port queues in the Q_FIFO and move the
9 associated chunks to an OUTPUT_DATA_FIFO. The OUTPUT_DATA_FIFO (not
10 shown) is located within reassembly block 205 and serves as the interface through
11 which data flows to outgoing SPI-4 interface block 206.

12 Outgoing SPI-4 interface block 206 periodically sends port full status information
13 that is used to load the "port full" register to keep reassembly block 205 updated on
14 how busy the various output logical ports are. Memory manager block 204
15 periodically sends empty status information that is used to load the "port empty"
16 register to keep reassembly block 205 updated on whether there is any more
17 information coming from the memory manager 205 for each of the various logical
18 output logical ports. In each of the "port empty" and "port full" registers, there is one
19 bit for each of the respective sixty-four logical output ports of ingress MS-SAR 125.

20 Reassembly block 205 steps through the rows of the port calendar of Figure 24,
21 one by one. For each row, reassembly block 205 reads the port ID (PID) from the
22 last field of the row and examines the bit corresponding to that port ID in each of the
23 "port empty" and "port full" registers. If the bits in these two registers indicate that
24 the port is neither full (i.e., that port has all the data it can handle) nor empty (no data
25 is available for that port), then reassembly block 205 pops the queue in Q_FIFO for
26 that port ID. The popping of the queue causes reassembly block 205 to retrieve one
27 64-byte data chunk from DATA_SRAM and its associated header from HDR_SRAM.
28 (In ingress mode, all chunks are 64 bytes but in egress mode EOP chunks can be
29 less than 64 bytes). The header is then appended to the data chunk as the complete
30 "switch cell" is pushed onto the OUTPUT_DATA_FIFO. Outgoing SPI-4 interface
31 block 206 pops the OUTPUT_DATA_FIFO, thereby causing the "switch cell" to be
32 supplied to outgoing SPI-4 interface block 206 in 64-bit pieces via 64-bit bus 325

1 (see Figure 10). The logical output port is provided to outgoing SPI-4 interface block
2 206 along with the switch cell. Outgoing SPI-4 interface block 206 in turn outputs
3 the switch cell onto SPI-4 bus 107 in 16-bit pieces via sixteen output terminals 326
4 of ingress MS-SAR 125. The SPI-4 bus 107 includes an SOP delimiter and an EOP
5 delimiter between data bursts on the SPI-4 bus. Data bursts on the SPI-4 bus are
6 transmitted in multiples of 16 bytes.

7 In this example, MS-SAR 125 is in the ingress mode, so there is only one logical
8 output port. Consequently there is only one entry in the port calendar of Figure 24.
9 When a row entry is processed having a jump bit of "1", processing of rows returns
10 to the first row. In the example of Figure 24 where there is only one logical output
11 port, there is a "1" in the jump column of the first row entry for port ID number one.
12 Accordingly, the one and only output queue in the Q_FIFO is popped. The circular
13 arrow in Figure 24 illustrates this return in the present example where there is only
14 one logical output port (PID 1).

15 Figure 27 illustrates information flow for flow #1 out of ingress MS-SAR 125. The
16 heavy upwardly pointing arrow represents the dequeuing operation being performed
17 by per flow queue block 207 including the passing of BIDs and associated
18 parameters to memory manager block 204. The heavy arrow extending from
19 payload memory block 217 to the right represents the flow of 64-byte chunks
20 through reassembly block 205 and to the switch fabric.

21 Per flow queue 207 continues issuing dequeue commands in the desired order
22 such that the associated various 64-byte chunks of the two flows (see Figure 23) are
23 read out of payload memory 217 and such that corresponding switch cells are
24 generated. The switch cells are output to switch fabric 105 from outgoing SPI-4
25 interface block 206 via SPI-4 bus 107.

26 Figure 28 illustrates the three switch cells that correspond to 64-byte chunks C1,
27 C2 and C3 of packet 300. The switch headers of each of the three cells contain the
28 "virtual port number" of line card 103 to which the cells are to be routed. The
29 Azanda header of each switch cell contains the egress application type (egress
30 application type 11) that the egress MS-SAR 200 is to perform on the switch cells of
31 flow #1.

1 Figure 29 illustrates the switch cell that corresponds to the one and only 64-byte
2 chunk for ATM cell 308. The switch header of this switch cell contains the "virtual
3 port number" of line card 103 to which the cell is to be routed. The Azanda header
4 of the switch cell contains the egress application type (egress application type 8) that
5 the egress MS-SAR 200 is to perform on the switch cell of flow #2.

6 Switch fabric 105 receives the switch cells from SPI-4 bus 107 and uses the
7 "virtual port number" embedded in the switch headers to route the switch cells to line
8 card 103. The switch cells are received on the SPI-4 bus 110 (see Fig. 9) of line
9 card 103. SPI-4 bus 110 is coupled to the SPI-4 bus input terminals (see input
10 terminals 198 of Figure 10) of egress MS-SAR device 200 of line card 103.
11 Operation of egress MS-SAR 200 is explained below in connection with Figure 10.

12 Figure 30 illustrates the general flow of information into egress MS-SAR 200.
13 The first incoming switch cell of packet 300 is read into incoming SPI-4 interface
14 block 201 and is passed to lookup block 202. Lookup block 202 need not, however,
15 perform a lookup using hashing as was done in the ingress mode described above.
16 Rather, because there is an Azanda header embedded in the switch cells, and
17 because that Azanda header contains the FID and the egress application type to be
18 performed by the egress MS-SAR, lookup block 202 merely locates this Azanda
19 header and extracts the embedded information.

20 Figure 31 illustrates the first switch cell for packet 300 of flow #1. The Azanda
21 header is the last four bytes of the switch header. Although the size of the switch
22 header can differ from switch fabric to switch fabric, all switch cells for a given switch
23 fabric have the same size. CPU 106 writes the size of the switch header for its
24 switch fabric into a "number of bytes of switch header" field in each row of the "port
25 table" (see Figure 14). This number is the same for all rows of the "port table".
26 Lookup block 202 in egress MS-SAR 200 reads the "number of bytes of switch
27 header" field in the row of the port table in the egress MS-SAR 200 for the logical
28 input port from which the switch cells were received. From this number of bytes of
29 the switch cell, lookup engine block 202 locates the last four bytes which are the
30 Azanda header. From the Azanda header, lookup block 202 retrieves the FID and
31 the egress application type to be performed by egress MS-SAR 200. Lookup block
32 202 removes the switch header and passes the remaining 64-bytes to segmentation

1 block 203 along with the FID and egress application type. The egress application
 2 type is supplied along with the 64-byte chunk so that the other blocks that receive
 3 the 64-byte chunk will perform the correct type of processing on the associated 64-
 4 bytes. The FID in the example of Figure 9 is flow #1. The egress application type in
 5 the example of Figure 9 for flow#1 is egress application type 11.

6 In accordance with egress application type 11, segmentation block 203 does not *perform* ✓
 7 segmenting per se but rather forwards the 64-byte chunk to memory manager block
 8 204 via the per-port FIFO mechanism described above. Memory manager block 204
 9 issues an enqueue command via enqueue command line 320, and stores the 64-
 10 byte chunk in payload memory 217. Per flow queue block 207 adds the BID to the
 11 per flow queue for flow #1. The same process flow occurs for the second and third
 12 switch cells associated with flow #1.

13 Figure 30 represents the storing of 64-byte chunks into payload memory 217 by
 14 the heavy upward pointing arrow that extends toward payload memory block 217.
 15 The heavy downward pointing arrow that extends toward per flow queue block 207
 16 represents the enqueue command and the building of the three BID linked list for
 17 flow #1.

18 The switch cell for the ATM cell (see Fig. 29) is received onto egress MS-SAR
 19 200 (see Fig. 9) in similar fashion. As in the case for the switch cells for the packet,
 20 lookup block 202 uses the logical input port number supplied by incoming SPI-4
 21 interface block 201 to access the appropriate row in the "port table" (see Fig. 14) in
 22 the lookup block 202. From the "number of bytes of switch header" field in that row,
 23 lookup block 202 locates the "Azanda header" in the incoming switch cell. The
 24 Azanda header contains the egress application type and the FID. Lookup block 202
 25 removes the switch header, and forwards the remaining 64-byte chunk to
 26 segmentation block 203 along with the FID and egress application type to be
 27 performed by egress MS-SAR 200. The 64-byte chunk in this case contains the 4-
 28 byte ATM header, the 48-byte ATM payload, and twelve bytes of pad.

29 In the example of Figure 9, the FID is flow #2. In the example of Figure 9, the
 30 egress application type for flow #2 is egress application type 8. In accordance with
 31 egress application type 8, segmentation block 203 does not perform segmentation
 32 per se but rather passes the 64-byte chunk and the egress application type to

1 memory manager block 204 via the per-port FIFO mechanism described above.
 2 Memory manager block 204 stores the 64-byte chunk. Per flow queue block 207
 3 adds the BID for this chunk and parameters associated with the chunk to a per flow
 4 queue for flow #2.

5 Once the 64-byte chunks from the various switch cells are stored in payload
 6 memory 217 and the linked lists for flow#1 and flow #2 are built, then the linked lists
 7 can be popped (i.e., "dequeued") and the 64-byte chunks output in various orders.
 8 For example, output scheduler block 210 can schedule the popping of the queues to
 9 control flow to output ports on a flow ID basis.

10 Figure 32 illustrates the general flow of information out of egress MS-SAR 200 in
 11 one scenario when the linked lists for flow #1 and flow #2 are dequeued. As each
 12 BID is dequeued from a per flow queue, the associated 64-byte chunk is read from
 13 payload memory 217 and is supplied via 128-bit~~e~~ wide data bus 324 to reassembly
 14 block 205. As explained above in connection with the ingress mode, reassembly
 15 block 205 maintains one reassembly queue for each of the 64 logical output ports. If
 16 two flows share the same logical output port, then the entire linked list for one flow
 17 must be dequeued before the linked list for the next flow having the same logical
 18 output port is dequeued. In the example of Figure 8, flow #1 and flow #2 have
 19 different logical output ports. The dequeue command as received by memory
 20 manager block 204 from per flow queue block 207 contains the port ID (PID) of the
 21 one of the 64 logical output ports of line card 103.

22 In the present example, the three-chunk linked list for flow #1 is dequeued first,
 23 one chunk at a time. Each 64-byte chunk is read into the DATA_SRAM of
 24 reassembly block 205 and a pointer to that chunk is stored in the Q_FIFO in the
 25 queue for the logical output port of flow #1. The first chunk contains the MPLS
 26 packet header. In an MPLS manipulation operation, reassembly block 205 can add
 27 (i.e., "push") MPLS labels to, or remove (i.e., "pop") MPLS labels from, the MPLS
 28 header found in the first chunk.

29 Consider the situation in Figure 9 where a first MPLS label for communication
 30 across the network coupled to router 100 via fiber optic cable 115 is to be replaced
 31 with a second MPLS label for communication across the network coupled to router
 32 100 via fiber optic cable 118. Reassembly block 205 determines the type of MPLS

1 manipulation to perform using the FID to lookup MPLS information that CPU 106
 2 stored previously in the external "header table" 327 (see Figure 10) of the egress
 3 MS-SAR 200. For each FID, header table 327 contains MPLS manipulation
 4 information.

5 Figure 33 is a more detailed diagram of an entry for a FID in "header table" 327 of
 6 Figure 10. The entry for each FID includes two 72-bit words. The first word contains
 7 an eight-bit control word CONTROL_WORD_1, a first 32-bit field that can contain
 8 either a first MPLS label or an ATM VPI/VCI header, and a second 32-bit field that
 9 can contain a second MPLS label. The second word contains a second eight-bit
 10 control word CONTROL_WORD_2, a first 32-bit field that can contain a switch
 11 header, and a second 32-bit field that holds the 32-bit Azanda header. The first two
 12 bits of CONTROL_WORD_1 indicate the header type: "00" indicates MPLS type,
 13 "10" indicates ATM type, and "01" indicates L2 Ethernet type. The next three bits of
 14 CONTROL_WORD_1 indicate the type of MPLS manipulation to perform: "000"
 15 indicates replace MPLS label and decrement the 8-bit MPLS time to live (TTL) field,
 16 "001" indicates push one MPLS label and decrement the original MPLS TTL, "010"
 17 indicates push two MPLS labels and decrement the original MPLS TTL, "011"
 18 indicates pop one MPLS label and decrement the original MPLS TTL, and "100"
 19 indicates pop two MPLS labels and decrement the original MPLS TTL. The next two
 20 bits of CONTROL_WORD_1 indicate the number of valid header words for that FID
 21 in the header table 327. The last bit of CONTROL_WORD_1 indicates that L2
 22 stripping and CRC checking is enabled. The first five bits of CONTROL_WORD_2
 23 indicate the number of L2 header bytes to remove if type is L2 Ethernet.

24 In the example of Figure 9, the entry in header table 327 for flow #1 indicates that
 25 the header type is MPLS and that information in the MPLS field (see Figure 33)
 26 determines the action. In the present example, the MPLS label is to be replaced.
 27 Reassembly block 205 therefore replaces the MPLS label in the first 64-byte chunk
 28 for flow #1 with the MPLS label found in an MPLS label field of the first word of the
 29 entry for flow #1 in header table 327. Reassembly block 205 also decrements the
 30 original TTL number.

31 As described above in connection with Figures 24-26, reassembly engine 205
 32 cycles through the entries in its port calendar (see Fig. 24). The port calendar is

1 provisioned beforehand by CPU 106 so that there is one entry for the port of flow #1
2 and another entry for the port of flow #2. If when reassembly block 205 services one
3 of the rows of the port calendar the full and empty registers indicate that outgoing
4 SPI-4 interface block 206 is ready for data for that output port, then the queue for
5 that output port is popped such that one 64-byte chunk stored in DATA_SRAM is
6 combined with a header from HDR_SRAM if appropriate and the result is pushed
7 onto the OUTPUT_DATA_FIFO within reassembly block 205.

8 As the various 64-byte chunks of the packet are received from memory manager
9 block 204, reassembly block 205 uses a CRC engine to calculate a CRC value on
10 the data. The CRC value calculated on only part of the data is called a "partial
11 CRC". The partial CRC value is changed as each additional piece of data passes
12 through the reassembly block 205. When the last of the data (as indicated by the
13 EOP bit) passes through the CRC engine, the CRC value is a "complete CRC". The
14 complete 32-bit CRC is compared with the 32-bit CRC value found in the AAL5
15 trailer of the last 64-byte chunk of the packet. If the two CRC values match, then the
16 data was not corrupted.

17 The AAL5 trailer also contains ^a sixteen-bit "length of payload" field that indicates
18 the length of the payload of the packet being reassembled. Reassembly block 205
19 maintains a partial packet byte count value for the packet. As each 64-byte chunk of
20 the packet is received from memory manager block 204, reassembly block 205 adds
21 the number of bytes received to the partial packet byte count value. After the last
22 64-byte chunk for the packet has been received and the partial packet byte count
23 value has been added to, reassembly block 205 compares the now complete packet
24 byte count value with the "length of payload" value from the AAL5 trailer. The two
25 should match.

26 After checking the CRC and the byte count, reassembly block 205 removes the
27 trailer and any padding from the last 64-byte chunk, and pushes the last chunk onto
28 the OUTPUT_DATA_FIFO.

29 Outgoing SPI-4 interface block 206 receives the pieces of the packet, piece by
30 piece (each piece being 64-bytes), from OUTPUT_DATA_FIFO and supplies those
31 pieces of the packet of flow #1 to framer 123. As framer 123 is ready to receive
32 another piece of the packet being sent out on a particular output port, outgoing SPI-4

1 interface block 206 sends it one via 16-bit wide SPI-4 bus 141. Outgoing SPI-4
 2 interface block 206 supplies bursts and start and end delimiters in accordance with
 3 SPI-4 protocol to demarcate the packet. Framer 142 receives the packet from SPI-4
 4 bus 141 and uses the start delimiters and the end delimiters to frame the packet.
 5 The packet passes through transmit serdes 143 and optics module 144 and onto
 6 fiber optic cable 118. In this way, 64-byte chunks of the packet of flow #1 go through
 7 reassembly block 205 one by one, the header of the first chunk being modified in
 8 accordance with egress application type 11, the trailer on the last chunk being
 9 checked and then removed. Accordingly, the term "reassembly" here does not
 10 necessarily require that the entire packet exist in reassembled form in the
 11 reassembly block or even on the MS-SAR. Rather the reassembly block outputs the
 12 various chunks of the packet sequentially such that the packet is transmitted in
 13 reassembled form on the fiber optic cable.

14 Figure 34 illustrates the three 64-byte chunks of flow #1 as passed through
 15 reassembly block 205 to outgoing SPI-4 interface block 206. The chunks are
 16 illustrated as 64-byte chunks rather than 64-bit pieces for ease of illustration.

17 Figure 35 illustrates the MPLS packet of flow #1 as output from SPI-4 interface
 18 block 206 to framer 142.

19 The 64-byte chunk for the ATM cell of flow #2 in this example is dequeued from
 20 payload memory 217 in similar fashion. Per flow queue block 207 issues a dequeue
 21 command to memory manager block 204 that indicates the BID of the chunk, and
 22 the logical output port from which the flow is to be transmitted. Reassembly block
 23 205 uses the FID of flow #2 to access the FID entry in "header table" 327 (see
 24 Figure 9). The header type bits of CONTROL_WORD_1 of the FID entry for flow #2
 25 indicates that the header type is ATM. Reassembly block 205 therefore replaces
 26 (i.e., "translates") the VPI/VCI field in the ATM header of the chunk in HDR_SRAM
 27 with the VPI/VCI information stored in the FID entry for flow #2. Reassembly block
 28 205 may also be programmed by information in header table 327 to change the CLP
 29 (Cell Loss Priority bit in the ATM header that indicates a high or low priority for the
 30 ATM cell).

31 Reassembly block 205 receives the 64-byte chunk, stores it into DATA_SRAM,
 32 and pushes a pointer to the 64-bytes onto a queue for the logical output port for flow

1 #2. Reassembly block 205 services the ports in its port calendar as explained
2 above, combines the data portion in the DATA_SRAM with the translated ATM
3 header portion in HDR_SRAM, and loads the completed ATM cell of flow #2 into the
4 OUTPUT_DATA_FIFO in reassembly block 205. The last eight bytes of the 64
5 bytes in DATA_SRAM are pad. Reassembly block 205 removes this 8-byte pad by
6 not reading it into the OUTPUT_DATA_FIFO. Outgoing SPI-4 interface block 205
7 pops the OUTPUT_DATA_FIFO such that reassembly block 205 supplies the
8 resulting 56-byte chunk to outgoing SPI-4 interface block 205. Figure 36 illustrates
9 the 56-byte chunk (ATM cell plus four bytes of pad) as output from reassembly block
10 205.

11 Outgoing SPI-4 interface block 206 removes the last four bytes of pad when
12 sending the ATM cell out. It does this using byte enables. Outgoing SPI-4 interface
13 block 206 outputs start and end delimiters to demarcate the 52-byte ATM cell in
14 accordance with the SPI-4 bus protocol. The entire 52-byte cell is output in one
15 burst. Figure 37 illustrates the ATM cell of flow #2 as output from outgoing SPI-4
16 interface block 206 of egress MS-SAR 200 via 16-bit wide SPI-4 bus 141.

17 Framer 142 receives the ATM cell from SPI-4 bus 141 and uses the start and end
18 delimiters to frame the ATM cell. Framer 142 then supplies the ATM cell to transmit
19 serdes 143. Serdes 143 in turn supplies the ATM cell in analog serial fashion to
20 optics module 144. Optics module 144 outputs the ATM cell onto fiber optic cable
21 118. In the example of Figure 9, there is only one physical output port (i.e., the fiber
22 optic cable 118). Framer 142 therefore maps the logical output ports for flow#1 and
23 flow #2 on SPI-4 bus 141 to the same serdes 143 for transmission on the same
24 wavelength channel on the same fiber optic cable 118.

25

26 EXAMPLE OF APPLICATION TYPES 5, 6, 14 AND 13:

27 Figure 38 illustrates an example of how two flows of different types (flow #1 is a
28 flow of three AAL5 cells, flow #2 is an MPLS packet) are received onto a first line
29 card (in this example, line card 101), simultaneously pass through ingress MS-SAR
30 125 on the first line card, pass through switch fabric 105 (in this example switch
31 fabric 105 is a packet-based switch fabric), pass onto a second line card (in this
32 example, line card 103), simultaneously pass through egress MS-SAR 200 on the

1 second line card 103, and are output by second line card 103. Flow #1 exits line
2 card 103 as an MPLS packet, whereas flow #2 exists line card 103 as four AAL5
3 cells. In this example, flow #1 and flow #2 are both communicated to first line card
4 101 on the same wavelength channel transmitted on the same fiber optic cable 115.
5 Similarly, flow #1 and flow #2 are both communicated from second line card 103 on
6 the same wavelength channel and on the same fiber optic cable 118.

7 In Figure 38, flow #1 is a flow of AAL5 cells flow passing into an ingress line card,
8 where the ingress line card supplies a packet to a packet-based switch fabric. As
9 indicated by the middle left example of Figure 8, this type of flow is identified as
10 ingress application type 5. The various blocks within ingress MS-SAR 125 therefore
11 perform processing on flow #1 in accordance with ingress application type 5. In
12 Figure 38, flow #2 involves an MPLS packet flow passing into an ingress line card,
13 where the ingress line card supplies a packet to a packet-based switch fabric. As
14 indicated by the bottom left example of Figure 8, this type of flow is identified as
15 ingress application type 6. The various blocks within ingress MS-SAR 125 therefore
16 perform processing on flow #2 in accordance with ingress application type 6.

17 Ingress application type 5 processing on flow #1 is explained in connection with
18 Figures 38, 5, 39 and 10. The first AAL5 cell of flow #1 is received via fiber optic
19 cable 115 (see Figure 5) and passes through optics module 119, serdes 121, framer
20 123 and classification engine 124. The 52-byte AAL5 cell is received onto ingress
21 MS-SAR 125 in 16-bit pieces via SPI-4 bus 135 via terminals 198 (see Fig. 10). The
22 AAL5 cell includes a 4-byte ATM header and a 48-byte payload. Incoming SPI-4
23 interface block 201 appends a 4-byte pad to the end of the 52-byte AAL5 cell before
24 forwarding the resulting 56-byte chunk 400 (see Fig. 39) to lookup block 202.
25 Figure 39 illustrates the 56-byte chunk 400 as it is output from incoming SPI-4
26 interface block 201.

27 As explained above in connection with the example of Figure 9, lookup block 202
28 uses the input port type supplied via bus 317 to lookup the associated traffic type in
29 the "port table" of lookup block 202. In this case, the traffic type is "ATM cells".
30 From the traffic type, lookup block 202 locates the VPI and VCI fields in the ATM cell
31 header. Lookup block 202 uses the VPI and VCI information to lookup in the "FID
32 table" (in external memories 215 and 216) the FID and the ingress application type.

1 Lookup block 202 forwards the FID and ingress application type to the other blocks
2 of the MS-SAR. Lookup block 202 removes the 4-byte ATM header and the 4-byte
3 pad, and forwards the remaining 48-byte chunk 401 of data to segmentation block
4 203 via bus 318. Figure 39 illustrates the 48-byte chunk 401 as it is output from
5 lookup block 202.

6 Segmentation block 203 adds a 16-byte pad to the 48-byte chunk to form a 64-
7 byte chunk 402. Memory manager block 204 reads the 64-byte chunk 402 from
8 segmentation block 203 and stores 64-byte chunk 402 into a buffer in payload
9 memory 217 using the enqueue command mechanism set forth above in connection
10 with the example of Figure 9. The per flow queue for flow #1 at this point involves
11 just one buffer. The other two AAL5 cells of flow #1 are processed in similar fashion
12 such that corresponding 48-byte chunks 403 and 404 are queued in the per flow
13 queue for flow #1.

14 Ingress application type 6 processing on flow #2 is explained in connection
15 with Figures 38, 5, 40 and 10. The MPLS packet flow #2 is received via fiber optic
16 cable 115 and passes through optics module 119, serdes 121, framer 123 and
17 classification engine 124. The MPLS packet is received onto ingress MS-SAR 125
18 in 16-bit pieces via SPI-4 bus 135 via terminals 198 (see Fig. 10). The MPLS packet
19 includes a four-byte MPLS header and a data payload. To illustrate operation of the
20 ingress MS-SAR, the MPLS packet data payload in this example is 150 bytes.

21 Incoming SPI-4 interface block 201 (see Fig. 10) forwards the 154-byte packet
22 405 as 64-bit pieces to lookup block 202. Figure 40 illustrates the 154-byte packet
23 405 as it is output from incoming SPI-4 interface block 201. As explained above in
24 connection with the example of Figure 9, lookup block 202 uses the input port type
25 supplied via bus 317 to lookup the associated traffic type in the lookup block's "port
26 table". In this case, the traffic type is "MPLS packet". From the traffic type, lookup
27 block 202 locates the 20-bit MPLS label within the MPLS header. Lookup block 202
28 uses the MPLS label to determine the FID and the ingress application type. Lookup
29 block 202 forwards the FID and ingress application type to the other blocks of the
30 MS-SAR. The packet 406 is forwarded in 64-bit pieces to segmentation block 203
31 via bus 318. Although all the 64-bit pieces of a packet for a given input port are
32 supplied to segmentation block 203 before any other 64-bit pieces are received for

1 that input port, 64-bit pieces for other input ports may be communicated from lookup
2 block 202 in an interleaved fashion with the 64-bit pieces of packet 406. Figure 40
3 illustrates the packet 406 as it is output from lookup block 202.

4 Segmentation block 203 "segments" the incoming packet 406 into 64-byte AAL5
5 like chunks. In the present example, there are three such chunks 407-409. The first
6 chunk 407 contains the 4-byte MPLS header. Segmentation block 203 adds an
7 AAL5 trailer 410 to the end of the last chunk 409. Any gap between the end of the
8 26-byte data payload 411 of the last chunk 409 and the trailer 410 is filled with a pad
9 412. The term "segmentation" here does not mean that segmentation block 203 in
10 this embodiment necessarily breaks the incoming packet into 64-byte chunks. That
11 has already been done in this embodiment by SPI-4 interface block 201. Rather, the
12 term "segmentation" means that the packet leaves segmentation block 203 properly
13 segmented into chunks in accordance with a segmentation scheme. In this
14 example, segmentation block 203 calculates a cyclic redundancy check (CRC) value
15 on the chunks of the packet and includes that CRC into trailer 410 of the last chunk
16 409.

17 Chunks of the packet are output from incoming SPI-4 interface block 201 in 64-
18 byte chunks. Between the various chunks of a packet, other chunks from other
19 flows may pass through the lookup block 202 and to the segmentation block 203.
20 Segmentation block 203 therefore queues the 64-byte chunks 407-409 into a FIFO
21 on a per port basis. The per port FIFO is located within segmentation block 203. As
22 segmentation block 203 receives each chunk from an input port, segmentation block
23 203 reads a partial CRC for the port from a CRC table (not shown) within
24 segmentation block 203. The partial CRC is modified as a function of the current
25 chunk received, and the new partial CRC is written back into the CRC table entry for
26 the appropriate port. When the EOP of a chunk indicates the chunk is the last chunk
27 of a packet, then the final CRC is calculated and this CRC is inserted into the trailer.

28 Memory manager block 204 reads the 64-byte chunks from the per port FIFO in
29 segmentation block 203 and stores the 64-byte chunks into buffers in payload
30 memory 217 using the enqueue command mechanism set forth above in connection
31 with the example of Figure 9. The per flow queue for flow #2 in this case involves
32 three buffers.

1 Once the 64-byte chunks for flow #1 and flow #2 are stored in payload memory
2 217, their per flow queues can be dequeued so as to carry out traffic shaping and/or
3 output scheduling functions. In the event that the per flow queue for flow #1 is
4 dequeued, 64-byte chunk 402 (see Figure 39) is output from memory manager block
5 204 via 128-bit bus 324 (see Figure 10). Reassembly block 205 uses the FID of
6 chunk 402 to lookup a switch header 413 in the "header table" 327. In this case,
7 CPU 106 has placed a switch header 413 appropriate for packet-based switch fabric
8 105 into the location for flow #1 in "header table" 327. Reassembly block 205
9 removes the 16-byte pad from the first chunk 402, adds the switch packet header
10 413 to the front to the chunk, and supplies the combination to outgoing SPI-4
11 interface block 206 using the port calendar mechanism explained above in
12 connection with the example of Figure 9. This outputting to outgoing SPI-4 interface
13 block 206 can occur before all the other 64-byte chunks of the per flow queue are
14 dequeued. In one embodiment of reassembly block 205, either the second 64-byte
15 chunk of a per flow queue or an EOP chunk must be received before the first chunk
16 can be output. The 64-byte chunks 403 and 404 are dequeued and processed by
17 reassembly block 205 in similar fashion, except that no header is added to these
18 chunks. The combination of the switch header 413 and the data portions of the
19 three chunks 402-404 is a switch packet 414.

20 Figure 39 shows this switch packet 414 as it is output from reassembly block 205
21 via 64-bit bus 325. The data portions of the three 64-byte chunks 402-404 together
22 form the data payload of switch packet 414. Figure 39 also illustrates switch packet
23 414 as it is output from outgoing SPI-4 interface block 206.

24 The 64-byte chunks 407-409 (see Figure 40) of flow #2 are dequeued in similar
25 fashion, the three chunks being processed through the per port queue in reassembly
26 block 205 for the output port of this flow. Reassembly block 205 uses the FID of the
27 first chunk to lookup a switch header 415 in "header table" 327. Reassembly block
28 205 adds switch header 415 to the front to the first 64-byte chunk. Reassembly
29 block 205 calculates a CRC from the data payload of the three chunks 407-409 and
30 checks to make sure that it matches the CRC found in the trailer 410 of the last
31 chunk 409. After checking the CRC, reassembly block 205 removes the pad 412
32 and trailer 410.

1 First chunk 407 is processed and supplied to the outgoing SPI-4 interface block
2 206 along with switch header 415 before the remaining chunks 408 and 409 of the
3 packet are dequeued. Reassembly block 205 does not store the entire packet in
4 reassembled form on the MS-SAR, but rather processes and outputs the individual
5 64-byte chunks of the packet one at a time. In this embodiment, the
6 OUTPUT_DATA_FIFO into which reassembly block 205 pushes processed chunks
7 is only 256 bytes in size, a size inadequate to store the entire 160-byte switch
8 packet.

9 Chunks of flow #2 are output from reassembly block 205 using the port calendar
10 mechanism explained above in connection with the example of Figure 9. The
11 combination of the data portions of the three chunks 407-409 and the added switch
12 header 415 together form switch packet 416. Figure 40 shows switch packet 416 as
13 it is output from reassembly block 205 via 64-bit bus 325. Switch packet 416 is
14 output from outgoing SPI-4 interface block 206 to the packet switch fabric 105 in 16-
15 bit pieces via terminals 326.

16 Switch packets 414 and 416 for flow #1 and for flow #2 are switched by the
17 packet switch fabric 105 of router 100 such that they are supplied to line card 103.
18 As described above in connection with the example of Figure 9, router 100 uses a
19 "virtual output port" number in the switch header to identify the particular destination
20 line card to which the packets should be routed.

21 Processing by egress MS-SAR 200 on flow #1 is explained in connection with
22 Figures 41, 5 and 10. Switch packet 414 from packet-based switch fabric 105 is
23 received in 16-bit pieces onto SPI-4 bus 100 via terminals 198. Switch packet 414
24 includes 8 or 16 bytes of switch header in addition to the packet itself. Figure 41
25 illustrates switch packet 414 as it is output from incoming SPI-4 interface block 201.
26 Switch packet 414 is passed to lookup block 202 along with a PID indicating the
27 logical input port.

28 In the example of Figure 38, flow #1 through egress MS-SAR 200 is a flow from a
29 packet-based switch fabric that is output as an MPLS packet. As indicated by the
30 flow to the bottom right of Figure 8, this is egress application type 14. The Azanda
31 header (see Figure 33) is the last four bytes of the switch header. This Azanda
32 header contains both the FID as well as the egress application type to be performed

on flow #1 by egress MS-SAR 200. Again, as explained above in connection with Figure 8, only one traffic type is permitted on any one logical input port. Lookup block 202 uses the logical input port number supplied to it via 64-bit bus 317 (see Figure 10) to locate the Azanda header in the information coming from the switch fabric. Lookup block 202 extracts the FID (FID1) and egress application type (egress application type 14). Lookup block 202 removes the switch header 413 and supplies the data payload of the packet via 64-bit bus 318 to segmentation block 203 along with the extracted FID and egress application type.

Figure 41 illustrates the data payload 417 as it is output from lookup block 202. Payload 417, however, is passed to segmentation block 203 in 64-byte chunks. In the example of Figure 38, the 144-byte payload of flow #1 contains enough information for there to be three such chunks. The first chunk 418 contains 64 bytes, the second chunk 419 contains 64 bytes, and the third chunk 420 contains sixteen bytes of data 421. Segmentation block 203 calculates a CRC on the data of the three chunks and adds a trailer 422 so that the trailer is located at the end of the third 64-byte chunk 420. Segmentation block 203 adds any necessary pad between the end of the data 421 of the last chunk and the trailer 422 of the last chunk. The three chunks 418-420 of Figure 41 are queued on a per port basis into an SRAM in segmentation block 203 as explained above in connection with Figure 40 and pointers to the chunks are pushed onto a FIFO in segmentation block 203. Memory manager block 204 pops the FIFO, reads the chunks, and stores the chunks into payload memory 217. Figure 41 illustrates the three 64-byte chunks 418-420 supplied by segmentation block 203 to memory manager block 204. The enqueue command mechanism as outlined above is used such that a per flow queue of BIDs is formed for the chunks of flow #1.

Processing by egress MS-SAR 200 on flow #2 is explained in connection with Figures 42, 5 and 10. The switch packet 416 from packet-based switch fabric 105 is received in 16-bit pieces onto SPI-4 bus 100 via terminals 198. Switch packet 416 includes 8 or 16 bytes of switch header 415, 4 bytes of MPLS header, and 150 bytes of MPLS data payload.

Figure 42 illustrates switch packet 416 as it is output from incoming SPI-4 interface block 201. Switch packet 416 is passed to lookup block 202 along with the

1 logical input port. The lookup block 202 uses the logical input port supplied to it via
2 64-bit bus 317 to locate the Azanda header in the information coming from the
3 switch fabric. The lookup block extracts the FID and egress application type (egress
4 application type 13). In the case of flow #2, information from a packet-based switch
5 fabric is output from egress MS-SAR 200 as AAL5 cells. This corresponds to egress
6 application type 13 as indicated in the middle right portion of Figure 8. Lookup block
7 202 removes switch header 415 and supplies the remainder of the packet (MPLS
8 header and MPLS payload) to segmentation block 203 along with the extracted FID
9 and egress application type. Figure 41 illustrates the MPLS header 423 and MPLS
10 data payload 424 as supplied by lookup block 202 to segmentation block 203 via 64-
11 bit bus 318.

12 In egress application type 13, segmentation block 203 "segments" the MPLS
13 header 423 and MPLS data payload 424 into four 48-byte chunks 425-428, because
14 each of these four chunks will be the data payload of an AAL5 cell to be output from
15 egress MS-SAR 200. In this example, the fourth 64-byte chunk contains only ten
16 bytes of data 429. Segmentation block 203 adds a trailer 430 so that the trailer 430
17 is located at the end of the first 48-bytes of the fourth 64-byte chunk. Any necessary
18 pad 431 is added between the end of data 429 and the start of trailer 430.
19 Segmentation block 203 adds a 16-byte pad to the end of each 48-byte chunk to pad
20 each chunk up to the 64-byte size stored in payload memory 217. Figure 42
21 illustrates the four 64-byte chunks as output from segmentation block 203. The
22 mechanism for supplying the four chunks to memory block 204 involves a queue as
23 explained above. Memory manager block 204 pops the queue, receives the 64-byte
24 chunks from segmentation block 203, and stores the 64-byte chunks into 64-byte
25 buffers using the enqueue command mechanism described above. A per flow
26 queue for flow #2 is created.

27 The pointers in the per flow queues for flow #1 and flow #2 are then popped off in
28 a desired order using the dequeue command. Here in this example of Figure 38, we
29 will consider the per flow queue of flow #1 being dequeued first. The three 64-byte
30 chunks for flow #1 are received from payload memory 217 by reassembly block 205
31 and are pushed onto a queue in reassembly block 205 for the intended output port.
32 The pad and trailer of the third 64-byte chunk 420 (see Fig. 41) are removed.

1 Reassembly block 205 also computes a CRC on the combined data of the three
2 chunks and verifies that the CRC in trailer 422 matches the newly computed CRC.
3 For the first 64-byte chunk 418, reassembly block 205 uses the FID to lookup an
4 MPLS header 432 in header table 327 (see Figure 10). The MPLS header 432 was
5 placed in header table 327 by CPU 106 before the lookup operation. Reassembly
6 block 205 performs an MPLS manipulation operation if controlled to do so by control
7 information in the header table.

8 Figure 41 illustrates a simplified view of MPLS header 432 and the aggregated
9 144 bytes of data 433 as it is output from reassembly block 205. MPLS header 432
10 and data 433 together form an MPLS packet 434. Although shown here
11 reassembled as a 144-byte block, the various 64-byte chunks of packet 434 are
12 output from reassembly block 205 one by one, the first 64-byte chunk having the
13 MPLS header appended to it. The chunks are output from reassembly block 205
14 using the port calendar and queue mechanism set forth above. Outgoing SPI-4
15 interface block 206 pops the Q_FIFO of reassembly block 205, receives the chunks
16 of the MPLS packet 434 in 64-bit pieces via 64-bit bus 325, and supplies the pieces
17 of the MPLS packet onto terminals 326. The pieces of MPLS packet 434 pass over
18 SPI-4 bus 141 to framer 142, through framer 142, serdes 143, and optics module
19 144, and onto fiber optic cable 118. It is therefore seen that ATM cell information
20 received onto the line card 101 passes through the packet-based switch fabric and
21 exits line card 103 as an assembled MPLS packet.

22 The per flow queue for flow #2 is dequeued next in this example (see Figure 42),
23 the 64-byte chunks 425-428 of flow #2 being supplied to the appropriate output port
24 queue in reassembly block 205. In egress application type 13, reassembly block
25 205 removes the 16-bytes of pad from each 64-byte chunk to recover the 48-bytes
26 of data. Reassembly block 205 uses the FID of a chunk to lookup a 4-byte ATM
27 header 435, adds the ATM header 435 to the 48-byte chunk of data, and adds a 4-
28 byte alignment pad such that each chunk is 56-bytes. The data 429, pad 431, and
29 trailer 430 of the fourth 64-byte chunk 428 form the data portion of the last 56-byte
30 chunk as illustrated in Figure 42. The ATM header 435 is the same for each of the
31 four 56-byte chunks of this flow.

1 Reassembly block 205 performs ATM translation by using the FID to access the
2 FID entry in "header table" 327 (see Fig. 10). Because the header type in
3 CONTROL_WORD_1 of the FID entry is "ATM", reassembly block 206 replaces the
4 PTI, CLP, EFCI and OAM bits in the ATM header using the incoming EOP, CLP,
5 EFCI and OAM bits from memory manager block 204. The last AAL5 cell (marked
6 EOP) is marked using the PTI field in the ATM header. The packet length is
7 checked against the maximum MTU. Reassembly block 205 checks the length and
8 CRC in AAL5 trailer 430. Figure 42 illustrates the AAL5 cells 436-439 as output
9 from reassembly block 205 in 64-bit pieces onto 64-bit bus 325. Outgoing SPI-4
10 interface block 206 pops the queue for the correct output port, retrieves each AAL5
11 cell, removes the 4-byte alignment pad, and outputs the resulting 52-byte AAL5 cells
12 onto terminals 326 in 16-bit pieces. The AAL5 cells of flow #2 pass over SPI-4 bus
13 141, through framer 142, through serdes 143, through optics module 144, and onto
14 fiber optic cable 118. It is therefore seen that MPLS packet information received
15 onto the line card 101 passes through the packet-based switch fabric and exits line
16 card 103 in AAL5 cell form.

17 APPLICATION TYPES 1 AND 9:

18 Figure 43 illustrates an example of ingress application type 1 and egress
19 application type 9. The incoming data in this example is two AAL5 cells. These two
20 52-byte AAL5 cells are received onto ingress MS-SAR 125 via terminals 198. The
21 two cells are output from the incoming SPI-4 interface block. The first cell contains 4
22 bytes of ATM header and 48 bytes of ATM data #1. The first part of the ATM data
23 #1 in this example is a packet header (for example, an IP header). The second cell
24 contains 4 bytes of ATM header and a small amount of data (ATM data #2). An
25 AAL5 trailer is disposed at the end of the 52 bytes of the second cell, and a pad is
26 disposed between the end of the AAL5 data #2 and the beginning of the trailer in
27 accordance with the AAL5 protocol.

28 Incoming SPI-4 interface block 201 adds four bytes of pad to each cell prior to
29 sending the cells to lookup block 202. The 4-byte pad is added because the data
30 path from the incoming SPI-4 interface block to the segmentation block is 64-bits
31 wide. Segmentation block 203 removes the 4-byte ATM header and adds an
32

1 additional twelve bytes of pad for a total of sixteen bytes of pad. Memory manager
2 block 204 passes the resulting two 64-byte chunks to payload memory 217 for
3 storage. The two 64-byte chunks are read from payload memory and are passed to
4 reassembly block 205. Reassembly block 205 does not in this ingress application
5 check the CRC in the trailer of the second cell. Reassembly block 205 adds either 8
6 or 16 bytes of a switch header to each 64-byte chunk, and passes the resulting
7 chunks to outgoing SPI-4 interface block 206 as illustrated. Chunks are, however,
8 processed through reassembly block 205 one at a time. As illustrated, the packet
9 header in the AAL5 data #1 is passed through the various blocks of the ingress MS-
10 SAR to the outgoing SPI-4 interface block 206. The cells pass out of the ingress
11 MS-SAR 125 as "switch cells", pass through the cell-based switch fabric, and to a
12 line card having an egress MS-SAR. The arrow from the top portion of Figure 43 to
13 the bottom portion of Figure 43 illustrates this passing through the switch fabric.

14 The two switch cells pass through the incoming SPI-4 interface block and to
15 lookup block 202 of the egress MS-SAR 200 as illustrated. Lookup block 202
16 removes the switch headers. The resulting two chunks pass through segmentation
17 block 203 and are passed to memory manager block 204 one at a time for storage.
18 Segmentation block 203 in egress application type 9 does not perform
19 segmentation. The two 64-byte chunks are read out of memory manager block 204
20 and pass to reassembly block 205. Reassembly block 205 maps AAL5 data #1 and
21 AAL5 data #2 into one contiguous data portion. As each of the data portions passes
22 into reassembly block 205, a CRC on the data is computed. The composite CRC is
23 then compared with the CRC in the trailer of the last 64-byte chunk (in this case the
24 second chunk) as that second chunk passes into reassembly block 205.
25 Reassembly block 205 adds an MPLS header and if programmed to do so by control
26 information in the header table, performs an MPLS manipulation operation on the
27 MPLS header. The resulting packet (MPLS header and AAL5 data #1 and AAL5
28 data #2) is then sent out to the framer via the outgoing SPI-4 interface block 206.

29

30 APPLICATION TYPES 2 AND 10:

31 Figure 44 illustrates an example of ingress application type 2 and egress
32 application type 10. In this example an incoming MPLS packet (MPLS header and

72 bytes of MPLS payload) is received onto the ingress MS-SAR 125 and passes through incoming SPI-4 interface block 201. Lookup block 202 performs a lookup and determines that the ingress application type is ingress application type 2. Segmentation block 203 outputs the packet as two 64-byte chunks such that the first 48 bytes of the first 64-byte chunk includes the MPLS header and a first part (AAL5 data #1) of the data payload. The remainder of the data of the MPLS packet (AAL5 data #2) is segmented into the second 64-byte chunk. Segmentation block 203 calculates a CRC on the 72 bytes of packet data and includes an AAL5 trailer into the second 64-byte chunk as illustrated so that the trailer is located at the end of the first 48 bytes of the second chunk. Segmentation block 203 appends 16 bytes of pad to the end of each 48-byte chunk so that each chunk is padded up to 64 bytes. The resulting two 64-byte chunks pass out of segmentation block 203 one at a time and are stored in memory manager block 204. The 64-bytes chunks are read out of payload memory and pass to reassembly block 205 one at a time. Reassembly block 205 performs a switch header lookup and adds a switch header to each 64-byte chunk. Reassembly block 205 does not perform any checking of the CRC in the trailer. The resulting switch cells are output via outgoing SPI-4 interface block 206 as illustrated.

The switch cells pass through a switch-based fabric and to an egress line card that includes the egress MS-SAR 200. The switch cells pass through the incoming SPI-4 interface block 201 of the egress MS-SAR 200. Lookup block 202 locates the Azanda header at the end of the switch header and determines the egress application type (egress application type 10) to be performed. Lookup block 202 removes the switch headers and passes the remaining two 64-byte chunks to the memory manager block 204 for storage into payload memory. The 64-byte chunks are read out of payload memory and are passed one at a time to reassembly block 205. Reassembly block 205 performs a header lookup and adds an ATM header to each 64-byte chunk. Reassembly block 205 removes 12 bytes of pad from the end of each chunk as illustrated in Figure 44. The resulting chunks are passed to outgoing SPI-4 interface block 206. Outgoing SPI-4 interface block 206 last the remaining 4-bytes of pad from the end of each chunk and outputs the chunks as ATM cells to the framer of the egress line card.

1

2 APPLICATION TYPES 4 AND 14:

3 Figure 45 illustrates an example of ingress application type 4 and egress
4 application type 14. In this example an incoming ATM cell (four bytes of ATM
5 header and 48 bytes of ATM payload) is received onto the ingress MS-SAR 125 and
6 passes through incoming SPI-4 interface block 201. Lookup block 202 performs a
7 lookup and determines that the ingress application type is ingress application type 4.
8 Lookup block 202 adds four bytes of pad as illustrated in Figure 45 and passes the
9 resulting 56 bytes to segmentation block 203. Segmentation block 203 generates
10 neither a CRC nor a trailer. Segmentation block 203 adds an additional 8 bytes of
11 pad as illustrated in Figure 45. The resulting 64-byte chunk is stored by memory
12 manager block 204 in payload memory. The 64-byte chunk is read out of payload
13 memory and passes to reassembly block 205. Reassembly block 205 performs a
14 header lookup and adds an 8 or 16-byte switch header to the 64-byte chunk. The
15 resulting switch packet (8 or 16 bytes of switch header, four bytes of ATM header,
16 48 bytes of ATM data, and 12 bytes of pad) is passed to a packet-based switch
17 fabric via outgoing SPI-4 interface block 206. The switch packet passes through the
18 switch fabric and is received onto the appropriate line card and the ingress MS-SAR
19 200. Lookup block 202 locates the Azanda header and determines the egress
20 application type (egress application type 14). Lookup block 202 removes the switch
21 header and passes the remaining 64-bytes to segmentation block 203.
22 Segmentation block 203 does not attach a trailer, but merely passes the 64-byte
23 chunk to memory manager block 204 for storage in payload memory. The 64-byte
24 chunk is retrieved from payload memory and is passed to reassembly block 205.
25 Reassembly block 205 performs a header lookup, and appends an MPLS header as
26 illustrated. Reassembly block 205 removes the last eight bytes of pad. The
27 resulting "encapsulated" ATM cell (MPLS header, 4 bytes of ATM header, 48 bytes
28 of ATM data, and 4 bytes of pad) is output as an MPLS packet to the framer of the
29 egress line card via outgoing SPI-4 interface block 206.

30

31 APPLICATION TYPES 6 AND 12:

1 Figure 46 illustrates an example of ingress application type 6 and egress
2 application type 12. In this example an incoming ATM encapsulated cell (four bytes
3 of MPLS header, 4 bytes of ATM header, and 48 bytes of ATM payload) is received
4 onto the ingress MS-SAR 125 and passes through incoming SPI-4 interface block
5 201. Lookup block 202 performs a lookup and determines that the ingress
6 application type is ingress application type 6. Lookup block 202 adds four bytes of
7 pad as illustrated in Figure 46 and passes the resulting 56-byte chunk to
8 segmentation block 203. Segmentation block 203 pads the chunk up to 64-bytes but
9 generates neither a CRC nor a trailer. The 64-byte chunk is stored by memory
10 manager block 204 in payload memory. The 64-byte chunk is read out of payload
11 memory and passes to reassembly block 205. Reassembly block 205 performs a
12 header lookup, adds an 8 or 16-byte switch header to the 64-byte chunk, and
13 removes the last 8 bytes of pad. The resulting packet (8 or 16 bytes of switch
14 header, four bytes of MPLS header, four bytes of ATM header, 48 bytes of ATM
15 data, and 4 bytes of pad) is passed to a packet-based switch fabric via outgoing SPI-
16 4 interface block 206. The switch packet passes through the packet-based switch
17 fabric and is received onto the appropriate line card and the egress MS-SAR 200. In
18 the egress MS-SAR, the lookup block 202 locates the Azanda header and
19 determines the egress application type (egress application type 12). Lookup block
20 202 removes the switch header and MPLS header and passes the remaining 56-
21 bytes (4 bytes of ATM header, 48 bytes of ATM data, and 4 bytes of pad) to
22 segmentation block 203. Segmentation block 203 does not attach a trailer, but adds
23 an additional 8-byte pad to pad the chunk up to 64 bytes. The 64-byte chunk is
24 stored in payload memory, is read out of payload memory, and passes to
25 reassembly block 205. Reassembly block 205 translates the ATM header. For
26 example, if instructed so by control information in the header table, reassembly block
27 205 replaces the VPI/VCI information in the ATM header with VPI/VCI information
28 retrieved from the header table. Reassembly block 205 removes the last 8 bytes of
29 pad and passes the resulting ATM cell (4 bytes of ATM header, 48 bytes of data,
30 and 4 bytes of pad) to outgoing SPI-4 interface block 206. Outgoing SPI-4 interface
31 block 206 removes the last 4 bytes of pad and passes the resulting 52-byte ATM cell
32 to the framer of the egress line card.

1

2 APPLICATION TYPES 6 AND 14:

3 Figure 47 illustrates an example of ingress application type 6 and egress
4 application type 14. In this example an incoming MPLS packet (four bytes of MPLS
5 header and 160 bytes of packet payload) is received onto the ingress MS-SAR 125
6 of a first line card and passes through incoming SPI-4 interface block 201. Lookup
7 block 202 performs a lookup operation and determines that the ingress application
8 type is ingress application type 6. Segmentation block 203 outputs the packet
9 information as 64-byte chunks as illustrated in Figure 47. In the example of Figure
10 47, the 160 bytes of data payload of the packet is segmented so that the 4-byte
11 MPLS header and 60 bytes of data is segmented into the first 64-byte chunk. The
12 next 64 bytes of data is segmented into the next 64-byte chunk. The remaining 36
13 bytes of data 440 from the 160-byte payload is segmented into the third 64-byte
14 chunk as illustrated. Segmentation block 203 calculates a CRC on the data payload
15 and places the CRC calculated into a trailer 441 that is incorporated as the end of
16 the last 64-byte chunk. Any space between the end of the data 440 and trailer 441
17 is filled with a pad 442. Segmentation block 203 may be outputting 64-byte chunks
18 of a packet at the same time that it is receiving chunks of the same packet. The 64-
19 byte chunks are stored by memory manager block 204 into payload memory. The
20 64-byte chunks are read out of payload memory and are supplied to reassembly
21 block 205. Reassembly block 205 maps the data portions of the 64-bytes chunks
22 into 160-bytes of packet data as illustrated. Reassembly block 205 calculates a
23 CRC on the data as the various chunks of data pass into the reassembly block 205.
24 When the data from the last 64-byte chunk is received onto reassembly block 205,
25 reassembly block compared the final CRC with the CRC in the trailer of the last 64-
26 byte chunk. Using the FID of the flow, reassembly block 205 performs a header
27 lookup and attaches a switch header to the front of the data payload. The resulting
28 switch packet (switch header, MPLS header, and packet data) is supplied to
29 outgoing SPI-4 interface block 206 in pieces, one at a time. Although the packet
30 data payload is illustrated in Figure 46 as being reassembled by reassembly block
31 205, a first part of the switch packet may be flowing out of reassembly block 205 to

1 outgoing SPI-4 interface block 206 at the same time that a second part of the data
2 portion of the switch packet is being received onto reassembly block 205.

3 The switch packet of Figure 46 passes through a packet-based switch fabric to a
4 second line card having an egress MS-SAR. The arrow in Figure 46 illustrates this
5 flow of information. The switch packet (switch header, MPLS header, and 160 bytes
6 of packet data in this example) is received onto the egress MS-SAR via incoming
7 SPI-4 interface block 201. Lookup block 202 removes the switch header, locates the
8 Azanda header, and determines from the Azanda header the egress application type
9 to be performed (egress application type 14). Segmentation block 203 segments the
10 packet as illustrated in Figure 46 into 64-byte chunks. Segmentation block adds a
11 trailer to the end of the last 64-byte chunk as explained above. The 64-byte chunks
12 are supplied one by one to the memory manager block 204 and are stored one by
13 one into payload memory. The 64-bytes are read out of payload memory one by
14 one and are supplied, one by one, to reassembly block 205. When reassembly
15 block 205 receives the first 64-byte chunk, reassembly block 205 performs a header
16 lookup using header table 327. If instructed to by control information found in the
17 header table 327, reassembly block 205 performs an MPLS manipulation operation
18 on the MPLS header. For example, reassembly block 205 may replace the MPLS
19 label in the MPLS header with an MPLS retrieved from header table 327.
20 Reassembly block 205 maps the data portions of the three 64-byte chunks into one
21 packet payload as illustrated in Figure 46. As the data portion of the various 64-byte
22 chunks pass through reassembly block 205, reassembly block 205 calculates a CRC
23 on the data such that reassembly block 205 checks the final composite CRC with the
24 CRC found in the last 64-byte chunk. The first part of the resulting MPLS packet can
25 be output from reassembly block 205 at the same time that later 64-byte chunks for
26 the packet are being received onto reassembly block 205. The resulting MPLS
27 packet (MPLS header, and 160 bytes of packet data) is output to the framer of the
28 egress line card via outgoing SPI-4 interface block 206.

29
30 REASSEMBLY "ON-THE-FLY":

31 Where a segmented packet is received on an egress line card in the form of a
32 plurality of individual cells (for example AAL5 cells), and where the data payloads of

1 the individual cells are to be reassembled into the original packet, the data payloads
2 of the individual cells are buffered in payload memory before any reassembly is
3 done. Rather than doing reassembly before the cells are buffered, reassembly in
4 the MS-SAR 200 is done on a per output port basis as the cells for a flow are being
5 output (i.e., on-the-fly) from the egress MS-SAR. A packet being reassembled is not
6 necessarily reassembled in the sense that the entire reassembled packet exists in
7 reassembled form in reassembly block 205. Rather the individual 64-byte chunks
8 making up the packet are processed by reassembly block 205 one at a time. It is,
9 for example, possible that a first part of the packet has already passed out of the
10 reassembly block 205 before the last part of the packet has been read out of
11 payload memory 204.

12 In accordance with one embodiment, a "reassembly context" is maintained for
13 each reassembly process going on. A reassembly context may, for example,
14 include: 1) a 32-bit partial CRC value maintained for a reassembly process taking
15 place on an output port, and 2) a 16-bit partial (i.e., running) packet length byte
16 count (count of the total number of bytes received for the reassembly process taking
17 place on the output port, including any padding bytes and any trailer bytes). Rather
18 than maintaining such a reassembly context for each of the many flows coming into
19 the egress MS-SAR (it is, for example, possible that each incoming flow would be
20 AAL5 cells requiring reassembly), reassembly is done on a per output port basis
21 after buffering in payload memory 217, thereby reducing the maximum number of
22 reassembly contexts that have to be maintained to one per output port. There is at
23 most one such packet reassembly process going on for each of the output ports that
24 is configured and active. Reducing the number of reassembly contexts to be stored
25 reduces the amount of memory necessary and thereby reduces line card costs.

26 27 SEGMENTATION "ON-THE-FLY":

28 A packet received onto an ingress line card may, in accordance with an
29 adaptation layer protocol, require segmentation into a plurality of pieces, where each
30 of the pieces is the data payload of an individual cell. The AAL5 protocol is an
31 example of one such protocol. The segmenting is done so that the individual cells
32 can later be reassembled into the original packet. Rather than storing such a packet

1 into payload memory, retrieving the packet, and then segmenting it as it is sent out
2 to the switch fabric, segmentation in ingress MS-SAR 125 is done on a per input port
3 basis as flows are received (i.e., on-the-fly) onto the ingress MS-SAR.
4 Segmentation, as the term is used here, encompasses the one-by-one processing of
5 segments of a packet such that the processed segments include information
6 required by the segmentation protocol. A segmentation block may, for example,
7 output a processed segment before the last part of the packet has even been
8 received by the segmentation block.

9 In accordance with one embodiment, a segmentation context is maintained for
10 each segmentation process going on. A segmentation context may, for example,
11 include a 32-bit CRC value and a 16-bit packet byte count value. Because a packet
12 to be segmented is received onto segmentation block 203 in segments, and
13 because segmentation block 203 processes these segments one by one, the
14 segmentation engine maintains a partial 32-bit CRC value for the packet and a
15 partial packet byte count value. As each segment is processed, segmentation block
16 203 updates the partial CRC value and partial packet byte count value. After the last
17 segment has been received onto the segmentation block as indicated by the EOP bit
18 of the segment, then the partial CRC value is the complete CRC value for the
19 packet. Similarly, the partial packet byte count value contains the number of bytes in
20 the packet. The 32-bit complete CRC value and the 16-bit packet byte count value
21 are included by segmentation block 203 into the AAL5-like trailer that the
22 segmentation block appends to the last 64-byte chunk for the packet. Rather than
23 maintaining such a segmentation context for each of the many flows going out of the
24 ingress MS-SAR (it is, for example, possible that each outgoing flow is to be in the
25 form of AAL5 cells), this segmentation process is done before buffering on a per
26 input port basis, thereby reducing the maximum number of segmentation contexts
27 that have to be maintained to one per input port. There is at most one such packet
28 segmentation process going on for each of the input ports that is configured and
29 active. Reducing the number of segmentation contexts to be stored reduces the
30 amount of memory necessary and thereby reduces line card costs. If a separate
31 counter or separate CRC engine is provided for each segmentation process, then

1 reducing the number of segmentation processes going on at one time further
2 reduces costs.

3
4 MULTIPLE DATA PATH CHIPS TO INCREASE THROUGHPUT RATE:

5 Figure 48 illustrates a novel aspect wherein MS-SAR functionality is partitioned
6 into a control integrated circuit and a data path integrated circuit such that system
7 throughput can be increased by using multiple data path integrated circuits. This
8 increase in system throughput is accomplished without having to redesign either the
9 data path integrated circuit or the control integrated circuit. The data path integrated
10 circuit and the control integrated circuit therefore make a versatile chip set. Where a
11 lower throughput is required, cost is reduced by using just one data path integrated
12 circuit with one control integrated circuit. Where a higher throughput is required,
13 multiple data path integrated circuits are used with one control integrated circuit.
14 Accordingly, the very same data path and control integrated circuits are useful for a
15 wide range of throughput applications. As a result, larger production runs of each
16 integrated circuit is likely and an associated reduction in per part cost is expected.

17 Figure 48 illustrates a line card 500 having a first MS-SAR 501 configured in the
18 ingress mode and a second MS-SAR 502 configured in the egress mode. Network
19 information passing from fiber optic cable 503 to the switch fabric passes through
20 optics module 504, serdes 505, framer/mapper 506 and classification engine 507
21 before it reaches ingress MS-SAR 501. In an egress path, network information
22 passes from egress MS-SAR 502, through framer/mapper 506, serdes 508 and
23 optics module 509 to fiber optic cable 510. The optics/serdes/framer circuitry of
24 Figure 48 is similar to the corresponding circuitry of Figure 5, except that the circuitry
25 in Figure 48 can receive from and output to fiber optic cables 503 and 510 are higher
26 data throughput rates. For example, where line card 101 of Figure 5 may receive
27 and output at OC-192 line rates (about 10 gigabits per second), line card 500 of
28 Figure 48 may receive and output at OC-768 line rates (about 40 gigabits per
29 second).

30 How the partitioning of MS-SAR functionality facilitates handling higher data
31 throughput rates is explained in connection with ingress MS-SAR 501 of Figure 48.
32 Ingress MS-SAR 501 includes a distribution integrated circuit 511, a control

1 integrated circuit 512, four data path integrated circuits 513-516, and one
2 aggregation integrated circuit 517. Control integrated circuit 512 includes the
3 circuitry shown in Figure 10 within dashed line 213. Each of the data path integrated
4 circuits 513-516 includes the circuitry shown in Figure 10 within dashed line 212.
5 Although per flow queue block 207 in the embodiment of Figure 10 has only one
6 control interface for sending and receiving enqueue and dequeue commands, the
7 control integrated circuit 512 of the embodiment of Figure 48 has four such
8 interfaces. Accordingly, ingress control integrated circuit 512 of Figure 48 is coupled
9 to data path integrated circuits 513-516 via control buses 518-521, respectively.

10 Operation of the circuit of Figure 48 is explained in connection with an example of
11 a high-speed stream of multiple packets being received from fiber optic cable 503 at
12 OC-768 rates. The packets pass, one by one, through optics module 504, serdes
13 505, framer/mapper 506, and classification engine 507 to distribution integrated
14 circuit 511. Distribution integrated circuit 511 receives the packets from
15 classification engine 507 via an SPI-4-like bus 522. This bus 522 is like an SPI-4
16 bus, except that it is capable of operating at 40 gigabits per second rates. Each
17 packet is framed by a start delimiter and an end delimiter.

18 Distribution integrated circuit 511 distributes each incoming packet to one of the
19 four data path integrated circuits 513-516, the particular data path integrated circuit
20 being chosen so as to distribute data path processing load evenly among the
21 multiple data path integrated circuits. Any of many known load-balancing algorithms
22 can be employed. Distribution integrated circuit 511 only outputs complete packets.
23 In addition to distributing load, distribution integrated circuit 511 includes a sequence
24 number along with each packet. In one embodiment, there is a sequence of
25 sequence numbers for each flow. Distribution integrated circuit 511 employs a
26 lookup block like lookup block 202 of Figure 10. Lookup information is provisioned
27 by the CPU beforehand as explained above in connection with lookup block 202 of
28 Figure 10. Distribution integrated circuit 511 receives a packet, identifies the flow,
29 increments a sequence number for that flow, and adds the sequence number to the
30 packet. Each respective packet of a flow therefore carries a sequence number that
31 is one greater than the sequence number of the previous packet in that flow.

FOI b7E b7C b7D

1 Figure 49 is a diagram of a packet as output by distribution integrated circuit 511.
2 The packet includes a data portion 523 and a header portion 524. The packet is
3 framed by a start delimiter 525 and an end delimited 526. The sequence number
4 527 added by distribution integrated circuit 511 is disposed between the start
5 delimited 525 and the packet header 524.

6 The various packets, after being marked with sequence numbers, flow into the
7 various data path integrated circuits 513-516. The lookup block of the data path
8 integrated circuit that receives the packet determines the flow ID of the packet as
9 described above in connection with the embodiment of Figure 10. In the
10 embodiment of Figure 48, the lookup block also extracts the packet sequence
11 number. Both the flow ID and the packet sequence number are then supplied to the
12 control integrated circuit 512 via the enqueue mechanism. Control integrated circuit
13 512 is therefore aware of which packets of which flows have been received onto
14 which data path integrated circuits. In addition to performing the other traffic
15 management, policing, shaping, and metering functions described above, control
16 integrated circuit 512 issues dequeue commands to ensure that the various packets
17 of a flow are supplied to aggregation integrated circuit 517 in the same order in
18 which they were received onto distribution integrated circuit 511. To do this, control
19 integrated circuit 512 maintains a "packet queue" for each flow.

20 Figures 50 and 51 illustrate the building of such a packet queue. In Figure 50,
21 the packets of a flow (FID1) were sequence numbered P1, P2, P3 and so forth by
22 distribution integrated circuit 511. Control integrated circuit 512 maintains a head
23 pointer and a tail pointer for this packet queue. The first packets P1, P2, P3 in this
24 flow have already been dequeued and sent through the aggregation integrated
25 circuit 517 to the switch fabric. Consequently they do not appear in the packet
26 queue. The next packet to be dequeued is packet P5. Control integrated circuit 512
27 will dequeue the various 64-byte chunks of this packet by dequeuing a per flow
28 queue for this packet as explained above in connection with Figure 10. In the
29 example of Figure 50, packet P7 of flow #1 has not been queued in the packet
30 queue. Accordingly, control integrated circuit 512 will not dequeue packet P8, but
31 rather will wait until packet P7 has been queued.

1 Figure 51 illustrates the packet queue after packet P7 has been queued. Note
2 that packet P7 has been linked into the queue in its proper place before packet P8
3 such that the dequeuing of the linked list will result in the packets of the flow being
4 dequeued in the proper order. In the example of Figure 51, packet P5 has been
5 dequeued, and a new packet for this flow, packet P10, has been received and
6 queued.

7 Control integrated circuit 512 dequeues the various packets in accordance with
8 this scheme so that the various packets of each flow are output to aggregation
9 integrated circuit 517 in the correct order. Aggregation integrated circuit 517
10 combines the packets in the order it receives them into one stream. That one
11 stream is output to the switch fabric via SPI-4-like bus 528. Bus 528 is like an SPI-4
12 bus, except that it is capable of operating at 40 gigabits per second rates. It is
13 therefore seen that one high throughput rate data path (40 gigabits/second) coming
14 into the line card is processed using four lower throughput rate data path integrated
15 circuits (10 gigabits/second), and that the outputs of the four lower throughput rate
16 data path integrated circuits are combined to form one higher throughput rate data
17 path (40 gigabits/second) to the switch fabric.

18 In the embodiment of Figure 48, the incoming data path having the increased
19 throughput rate is being controlled by only one control integrated circuit 512. The
20 associated increase in processing required of control integrated circuit 512 may
21 result in difficulties in accessing external memory.

22 Figure 52 is a diagram of an external memory device 529 that is coupled to a
23 control integrated circuit. External memory device 529 may, for example, be
24 external memory 225 of Figure 10. External memory device 529 stores two types of
25 information, information #1 and information #2, both of which must be accessed
26 within a particular amount of time related to the rate of incoming information. Where
27 the control integrated circuit is clocked by a clock signal, this particular amount of
28 time can be referred to as eight clock periods. In the example to the left of Figure
29 52, each piece of information can be accessed in two clock periods. Both pieces of
30 information are stored in the same external memory device requiring one to be
31 accessed before the other. A total of four clock periods is therefore required to
32 access both pieces of information.

1 If the circuit of Figure 48 is now employed to handle OC-768 line rate information
2 coming in from fiber optic cable 503, then this same information (information #1 and
3 information #2) must be accessed in a smaller amount of time. Consider the
4 example where both information #1 and information #2 must now be accessed within
5 two clock periods. A memory may not be available that has fast enough access
6 times to handle the accessing required within this fewer number of clock periods.

7 In accordance with one novel aspect, two external memories are used.
8 Information #1 is stored in the first external memory 530 and information #2 is stored
9 in second external memory 531. The two external memories are accessed at the
10 same time in parallel. If external memories 530 and 531 are the same type of
11 external memory 529, then these external memories have access times of two clock
12 periods and both information #1 and information #2 are accessed within the required
13 two clock periods. It is to be understood, of course, that the eight and two in the
14 example are used only to illustrate the technique of accessing memories in parallel
15 to facilitate handling higher data throughput rates. The technique here is not limited
16 to the numbers in this example. An example of information #1 is cell count
17 information stored in the embodiment of Figure 10 in PFQ STAT memory 225. An
18 example of information #2 is packet count information stored in the embodiment of
19 Figure 10 in PFQ STAT memory 225. These two types of information are, in one
20 embodiment of Figure 48, stored in different external memory devices.

21 Although the operation of the embodiment of Figure 48 is described in
22 connection with a flow of packets, the same process is followed to handle flows of
23 cells. The embodiment of Figure 48 handles all the application types of Figures 7
24 and 8. It is also to be understood that the functionality of the circuit of Figure 48 can
25 be rearranged. For example, the distribution chip and the control integrated circuit
26 may be combined onto one integrated circuit such that only one lookup block is
27 required. Alternatively, the lookup function may be performed on the distribution
28 chip such that results of the lookup are forwarded to the control integrated circuit.
29 Various rearrangements of the functionality of Figure 48 are possible without losing
30 the benefit of the novel partitioning of the MS-SAR.

31
32 BACKPRESSING USING SERIAL BUS:

1 A router 600 involves a first line card 601 and a second line card 601. Each of
2 the first and second line cards involves an MS-SAR operating in the ingress mode
3 and an MS-SAR operating in the egress mode. The egress MS-SAR 603 on the
4 second line card can become endangered of being overloaded if, for example, the
5 ingress MS-SAR 604 on the first line card continues to send network information for
6 a flow to the egress MS-SAR 603 on the second line card, but the egress MS-SAR
7 603 on the second line card is prevented from outputting that information, for
8 example due to congestion at the framer 605. Consequently, more and more of the
9 network information for the flow will end up having to be buffered by the egress MS-
10 SAR 603 of the second line card (buffered in payload memory).

11 In one novel aspect, the ingress and egress MS-SAR devices 604 and 606 of the
12 first line card 601 are linked by a serial bus 607 on the first line card, and the ingress
13 and egress MS-SAR devices 603 and 608 of the second line card 602 are linked by
14 a serial bus 609 on the second line card. If the egress MS-SAR 603 of the second
15 line card is in danger of becoming overloaded, then the egress MS-SAR 603 of the
16 second line card sends an indication of this situation to the ingress MS-SAR 608 of
17 the second line card via the serial bus 609 on the second line card. In one
18 embodiment, it is the per flow queue block of the egress MS-SAR 603 that detects
19 that the size of the free buffer queue has decreased to an undesirably low level. The
20 per flow queue block is therefore coupled to the serial bus 609 as illustrated in
21 Figure 53.

22 The ingress MS-SAR 608 of the second line card receives this indication from
23 serial bus 609 and relays the indication to the first line card 601 by outputting a
24 special status switch cell 611. The special status switch cell 611 is transported
25 across the switch fabric 610 to the egress MS-SAR 606 of the first line card. The
26 egress MS-SAR 606 of the first line card detects the special status switch cell, and
27 relays the indication of the situation to the ingress MS-SAR 604 of the first line card
28 via the serial bus 607 on the first line card. In one embodiment, it is the
29 segmentation block that detects the special status switch cell. The segmentation
30 block is therefore coupled to the serial bus 607 as illustrated in Figure 53.

31 In response to receiving this indication from serial bus 607, the ingress MS-SAR
32 604 on the first line card slows or stops its outputting of the information that is

1 overburdening the egress MS-SAR 603 on the second line card. In one
2 embodiment, the output scheduler is able to slow or stop the outputting of
3 information that is overburdening egress MS-SAR 603. Consequently, the serial bus
4 607 is coupled to the output scheduler block of ingress MS-SAR 604 as illustrated in
5 Figure 53.

6 7 INCOMING SPI-4 INTERFACE BLOCK IN MORE DETAIL:

8 Figure 54 is a block diagram of one particular embodiment 800 of incoming SPI-4
9 interface block 201 of Figure 10. Incoming SPI-4 interface block 800 includes an
10 input control block 801, an input FIFO block 802, an output control block 803, a port
11 calendar block 804, a CPU interface block 805, and a test multiplexer block 806. As
12 illustrated in Figure 55, input control block 801 includes a single-data rate (SDR)
13 pack block 807, a receive (RX) state machine block 808, and a parity checker block
14 809. As illustrated in Figure 56, output control block 803 includes a per port control
15 block 810, an SPI table SRAM block 811, a data FIFO 812, a control FIFO 813, an
16 output control block 814, and a free buffer list block 815.

17 In operation, incoming SPI-4 interface block 800 receives 16-bit double data rate
18 data at approximately 10 gigabits per second that is transmitted to the MS-SAR in
19 accordance with the SPI-4, phase II, specification. Incoming SPI-4 interface block
20 800 packs the incoming data into 64-bit wide words, extracts in-band port control
21 information (input port addresses, SOP, EOP, and error control codes), and then
22 passes the data along with the extracted control information to lookup block 202
23 (see Figure 10). SDR pack block 807 converts the 16-bit DDR data words into 32-bit
24 SDR data words. The one-bit DDR control bit is converted into a two-bit SDR
25 control word. The two-bit control word indicates which 16-bit word within a 32-bit
26 word is the control word. Based on the two-bit control word, RX state machine 808
27 extracts the in-band port address, start/end-of-packet indication and error-control
28 code from the data and control path and packs the data into 64-bit data words and
29 pushes the 64-bit data and the 16-bit control into 80-bit wide input FIFO block 802.

30 Per-port control block 810 of Figure 56 accumulates data on a per port basis and
31 queues the control information into control FIFO 813 when it has accumulated a
32 complete data buffer in data FIFO 812. Output control block 814 reads from control

1 FIFO 813 and from data FIFO 812 and outputs the related data and control
2 information to lookup block 202. As the data flows into per port control block 810,
3 parity checker block 809 calculates four-bit Diagonal Interleaved Parity (DIP-4) and
4 reports an error if the calculated DIP-4 is not equal to the one received with the
5 control word.

6 More detailed operation of the various blocks of Figures 54-56 is now described.
7 SDR pack block 807 of Figure 55 converts the 16-bit 400 MHz DDR data into 32-bit
8 400 MHz SDR data. The RDCTL control signal is used to determine whether the
9 high byte or the low byte of the 32-bit SDR data is control information. The RX state
10 machine block 808 of Figure 55 receives 32-bit words from SDR pack block 807
11 along with a two-bit indication of data or control. RX state machine 808 extracts the
12 in-band control information and packs the data into 64-bit wide words. It also sends
13 the data and control words to parity checker block 809 to check parity.

14 The incoming SPI-4 interface block 800 expects the control word be sixteen bits
15 and to have a particular format. Bits 0-3 are a DIP-4 (Diagonal Interleaved Parity)
16 value. Bits 4-11 are an 8-bit port address. Bit 12 is a start-of-packet (SOP) bit. Bits
17 13 and 14 are end-of-packet status bits. Bit 15 is a control word type. Such a 16-bit
18 control word, when present in the data path, is aligned such that its MSB is sent on
19 the MSB of the incoming data lines. A control word that separates two adjacent data
20 burst transfers may contain status information pertaining to the previous transfer, the
21 following transfer, or both transfers.

22 RX state machine block 808 operates in accordance with a state transition table,
23 where transitions are based on the types of control words received. Parity checker
24 block 809 calculates DIP-4 parity and compares the calculated parity with the one
25 received in the control word.

26 Input FIFO block 802 (see Figure 54) is a 256x80 bit dual-port FIFO running at
27 200 MHz that absorbs the incoming data and control information from input control
28 block 801. Each 80-bit entry includes an SOP bit, an EOP bit, an SOB bit, an EOB
29 bit, an error bit, a parity error bit, on reserved bit, five PID bits, three byte valid bits,
30 and 64 data bits. The write operation is controlled by receive state machine 808.
31 Read operation is controlled by the per-port control block 810 (see Figure 56). Per-

1 port control block 810 reads from input FIFO block 802 as long as input FIFO block
2 802 is not empty.

3 For each new data burst read from FIFO 802 (a burst of data is marked with SOP
4 or SOB), per-port control block 810 obtains a pointer to a free buffer in data FIFO
5 812 from the free buffer list block 815. Free buffer list block 815 includes a 128x10
6 bit single port SRAM containing pointers to available locations in data FIFO 812.
7 Per-port control block 810 also obtains temporary control information from SPI table
8 block 811. SPI table block 811 is a table of sixty-four entries, one for each input
9 port. The table contains control information for the incoming data burst for each
10 port. Each 32-bit entry in SPI table 811 includes one SOP bit, one EOP bit, one
11 error bit, one parity error bit, six PID bits, three byte valid bits, seven cell count bits,
12 ten pointer bits, and two reserved bits.

13 After obtaining a pointer to a free buffer and after obtaining associated control
14 information from SPI table block 811, per-port control block 810 accumulates the
15 incoming data into the pointed to 80-byte buffer in data FIFO 812 (accumulates on a
16 per-port basis, one buffer per port). The associated control information is only
17 pushed into the control FIFO 813 when the buffer is full or when an EOP is received.

18 Data FIFO 812 is a dual-port FIFO that implements 128 buffers of 80 bytes each.
19 The amount of data stored in a buffer is sixty-four bytes in ingress mode, and eighty
20 bytes in egress mode when operating with a packet-based switch fabric. When
21 operating in egress mode with a cell-based switch fabric, each incoming switch cell
22 is a complete packet (i.e., there is an EOP for each incoming switch cell). Each
23 buffer is used to store either seventy-two bytes or eighty bytes, depending on
24 whether the switch header is eight or sixteen bytes long. When data from an 80-
25 byte buffer is sent out to lookup block 202, output control block 814 returns the
26 pointer to the buffer back to the end of the free buffer list stored in free buffer list
27 block 815.

28 Port calendar block 804 receives per-port FIFO status information from
29 segmentation block 203 via line SEG_SPII_FULLL. Port calendar block 804 encodes
30 this status as a two-bit status word and sends the status word back to the transmit
31 device at the rate of one port per cycle in round robin fashion based on the port
32 calendar. The "11" status word is reserved for in-band framing. A DIP-2 odd parity is

1 There is one 74-bit entry in control queue FIFO 912 for each of the 64-byte buffers in
2 data SRAM block 908. Consequently, there are eighty 74-bit entries in control
3 queue FIFO 912. Control queue FIFO is a dual-port SRAM. As bursts of data flow
4 into segmentation block 900 from lookup block 202 (see Figure 10), the incoming
5 data is stored in data SRAM block 908 whereas control information for the data is
6 stored in control queue FIFO 912. There is a one-to-one correspondence between
7 the 64-byte buffers in data SRAM 908 and the entries in control queue FIFO 912.
8 Free buffer list block 905 stores a pointer that points to one pair of a 64-byte buffer in
9 data SRAM 908 and its associated 74-bit entry in control queue FIFO 912. The
10 pointer indicates whether the buffer is available or not.

11 One 74-bit entry in control queue FIFO 912 includes: 1) seven bits for the port ID,
12 2) four bits for the type of incoming traffic on the port, 3) one SOP bit which if set
13 indicates the start of packet, 4) one EOP bit which if set indicates the end of packet,
14 5) one BAD bit which if set indicates an error condition and that the data should be
15 discarded, 6) one CLP (Cell Loss Priority) bit which is valid only for ATM traffic, 7)
16 one EFCI (Explicit Forced Congestion Indication) bit which is valid only for ATM
17 traffic, 8) one OAM (Operation, Administration and Maintenance) bit, 9) one LAST
18 AAL5 bit which if set indicates the last chunk of an AAL5 group of chunks, 10) one
19 AAL5 TRAILER bit which indicates that the data in the current buffer has an
20 associated trailer, 11) six "cell length" bits (also called "chunk length" bits) which
21 indicate the number of valid bytes in the 64-byte chunk, 12) seven PTR bits which
22 indicate the start address of the buffer in data SRAM 908, 13) three reserved bits,
23 14) twenty FID bits which indicate the flow ID of the incoming data on the port, 15)
24 sixteen "packet length" bits which indicate number of bytes in the packet (only valid
25 for packet traffic), and 16) three QOS (Quality Of Service) bits that indicate a quality
26 of service.

27 Segmentation table and statistics SRAM block 907 actually includes two dual-port
28 memories, a segmentation table memory and a statistics memory. For each input
29 port, there is a 36-bit entry in the segmentation table memory that contains per port
30 parameters. Each segmentation table entry includes: 1) one SOP bit which if set
31 indicates a start of packet burst has been received for the port, 2) one EOP bit which
32 if set indicates that an end of packet burst has been received for the port, 3) one

1 BAD bit which if set indicates an error condition has been detected and that the
2 associated data should be discarded, 4) seven PTR bits which indicate the start
3 address of the associated 64-byte chunk in data SRAM 908, 5) the sixteen "packet
4 length" bits which are used to keep track of the total number of bytes of the incoming
5 packet that have been received on the port, 6) the six "cell length" bits (also called
6 "chunk length" bits) which are used to keep track of the total number of bytes
7 received for the associated 64-byte buffer in data SRAM 908, and 7) four reserved
8 bits. The "cell length" value indicates how big the chunk is (in bytes) that is present
9 in the associated 64-byte buffer. There are sixty-four such 36-bit entries, one for
10 each of the sixty-four input ports.

11 For each input port, there is also an 80-bit entry in the statistics memory. Each
12 statistics table entry includes: 1) a 32-bit "packet received" value which indicates the
13 number of packets received on that port, and 2) a 48-bit "data received" value which
14 indicates the number of bytes of data received on that port. There are sixty-four
15 such entries, one for each input port.

16 Information coming into the MS-SAR is output from the incoming SPI interface
17 block 201 in bursts. Each burst is a multiple of sixteen bytes. Some bursts can be
18 longer than others. Each burst is received associated with one of the sixty-four
19 possible input ports. The incoming SPI interface block 201 adds a start of burst
20 (SOB) bit to the beginning of the burst, and adds an end of burst (EOB) bit to the
21 end of the burst. The data of an MPLS packet or an ATM cell may be carried in
22 multiple such bursts. Segmentation block 900 receives these bursts from lookup
23 block 202. A burst from one input port can be followed by a burst from a different
24 input port such that multiple bursts that comprise the data for one packet are
25 received interspersed with intervening bursts for other packets.

26 In normal operation, if there is a 64-byte buffer available in data SRAM 908 to
27 receive an incoming burst, then free buffer list block 905 outputs a seg_lut_rdy
28 signal 918 to lookup block 202. Input phase state machine 904 waits for a
29 lut_seg_valid signal 919 to be returned from the lookup block 202. The returned
30 lut_seg_valid signal 919 indicates that there is valid control information and data
31 information on the 64-bit bus 318 from lookup block 202. In this example, the
32 lut_seg_valid signal 919 being asserted indicates that all signals whose signal

1 names start with lut_seg are valid. Sixty-four bits of data are clocked in at once from
2 lines lut_seg_data[63:0], the number of valid bytes in the sixty-four bits being
3 indicated by lut_seg_byte_v[2:0].

4 As information is received from a given input port, the values in the segmentation
5 table entry associated with the input port and values in the statistics table associated
6 with the input port are updated. For example, as explained above, the segmentation
7 table includes a six-bit "cell length" (also called "chunk length") field for storing the
8 number of bytes in the chunk of data that is stored in the current 64-byte buffer in
9 data SRAM 908. Because the data from multiple bursts can be placed into the
10 current 64-byte buffer in data SRAM 908, and because traffic from other ports can
11 be received onto the segmentation block 900 between these multiple bursts,
12 segmentation block 900 keeps track of the segmentation table information and the
13 statistics information between bursts. Accordingly, the first time a burst is received
14 for a 64-byte buffer for a particular input port, input phase state machine 904 writes
15 initial information into the entry in segmentation table memory. The SOP bit, for
16 example, is set whereas the EOP bit is not. The "packet length" field and the "cell
17 length" field are set to zero. If, on the other hand, a burst has already been received
18 for that particular input port, then input phase state machine 904 reads the per-port
19 parameters from the entries in the segmentation table memory 907 for that input
20 port. Similarly, input phase state machine 904 reads the statistics in the statistics
21 memory 907 for that input port. Input phase state machine 904 reads these values
22 when the lut_seg_valid signal 919 is received from lookup block 202. More
23 particularly, when the lut_seg_valid signal 919 is received from the lookup block 202,
24 the input phase state machine 904: a) reads from the free buffer list SRAM block
25 905 to obtain a pointer to an available 64-byte buffer in data SRAM block 906, b)
26 uses the port ID received from lookup block 202 to read segmentation table SRAM
27 907 to get the per-port parameters for that port, and c) uses the port ID received
28 from lookup block 202 to read the statistics SRAM 907 to get the statistics for that
29 port. The per-port parameters received from lookup block 202 with the burst are
30 compared with those from segmentation table memory 907. Error conditions may be
31 set. For example, if two consecutive SOP bursts are received without an intervening
32 EOP burst, then an error has occurred. If an error condition is detected, then the

1 BAD bit in the segmentation table SRAM entry is set to indicate that the associated
2 data is bad and should be discarded.

3 In Figure 57, input phase state machine 904 contains a signal labeled cur_xxx.
4 This label represents many signals associated with the current burst. The label
5 prv_xxx represents many signals associated with the previous burst. One of the
6 current signals cur_xxx is the port ID associated with the current burst. This current
7 port ID signal indicates the last port address processed. The port ID received from
8 the bus 318 is compared to the current PID to see if the incoming burst is from the
9 same port address. If so, then the data burst is a continuation to the one previously
10 processed, and segmentation table and statistics information already in internal
11 registers in input phase state machine 904 are used. The segmentation table and
12 statistics SRAM 907 does not need to be accessed. If, on the other hand, the
13 incoming burst is not from the same port, then the incoming data burst is for a new
14 port ID and new segmentation table and statistics information is to be used.
15 Segmentation table and statistics information for the last processed burst are
16 therefore stored back into the segmentation table and statistics memory 907 before
17 similar information for port of the new burst is loaded into the registers in the input
18 phase state machine 904. There are three sets of internal registers (not shown) in
19 input phase state machine 904 that store three sets of per-port segmentation table
20 and statistics information: one for the last processed port (i.e. the current
21 parameters); one for the port from lookup block 202 that comes with the valid signal;
22 and one for the previously serviced port that needs to be written back to the
23 segmentation and statistics memory 907.

24 There are two statistics counters in the input phase state machine 904, one that
25 counts the number of packets received on the current port (32-bit "packet received"
26 value) and one that counts the number of bytes received on the current port (48-bit
27 "data received" value). The outputs of these counters are kept in the statistics
28 memory 907 on a per port basis. The running count of bytes received and the
29 number of packets received continue to accumulate and may roll over. The CPU of
30 the router can read the two values at two different times, take the differences
31 between the values at the two times, and thereby calculate how much traffic is being
32 received on each port. The CPU reads these values via the CPU interface 913.

1 There is at most one segmentation process going on per port. Incoming burst
2 data for a port is accumulated into a corresponding one of the 64-byte buffers in data
3 SRAM 908. The maximum amount of data that can be stored in the 64-byte bluffer
4 (either 48, 56, or 64 bytes) is determined by the application type of the flow. The 64-
5 byte format is used for packet data. The 48-byte format is used for ATM data. The
6 56-byte format is used for other ATM data. Control information for the data in a 64-
7 byte buffer is queued into a location in queue FIFO 912 that corresponds to the
8 particular 64-byte buffer. Segmentation table and statistics information is written
9 back to the segmentation table and statistics memory 907. Input phase state
10 machine 904 contains a six-bit data byte counter (not shown) that increments the
11 six-bit "cell length" (also called the "chunk length") count value (stored in
12 segmentation table memory 907) upon receiving each incoming data byte for a 64-
13 byte buffer. When the number of bytes in the 64-byte buffer reaches the maximum
14 amount allowed by the format used, then the data in the 64-byte buffer is ready for
15 transfer to memory manager block 204.

16 Input phase state machine block 904 also contains a 16-bit packet length counter
17 that counts the number bytes in the current packet. The output of this packet length
18 counter is the "packet length" field in the queue FIFO entry. This counter continues
19 counting bytes until the EOP is reached. The output of this counter is incorporated
20 into the AAL5 trailer if such a trailer is to be added.

21 Segmentation block 900 handles the alignment of data to 64-bit bus 318 (see
22 Figure 10) in the situation where lookup block 202 adds a L2 special header that
23 does not line up with a 64-bit boundary. Align pipe block 903 is a pipeline used to
24 align control information (including port ID (PID) and flow ID (FID)) so that the control
25 information lines up with the 64-bits of data passing through the L2 align block 901.
26 Block 902 is not used to calculate a CRC for an AAL5 trailer. Rather, block 902 is
27 used only to calculate the CRC-32 for an L2 header in the special case where a
28 packet is being broken up into multiple AAL5 cells that are being output in the egress
29 mode. In such a case, the CRC bytes in the L2 header are replaced by new CRC
30 bytes calculated by L2 CRC generator 902. If no L2 header modification is to be
31 performed, then the data passes to the input phase state machine 904 via a bypass
32 path such that no alignment is performed.

Free Buffer List block 905 is an 80x7 bit dual port internal SRAM. After power on reset, the CPU initializes the free buffer list using a CPU command called the "Init FBL" command. There are multiple CPU commands. The CPU can cause a CPU command to be performed writing a command opcode for a particular CPU command into a command register in CPU interface block 913. CPU interface block 913 involves a state machine that then carries out the command indicated by the opcode. There are commands by which the CPU can write information into any location in any of the memories within segmentation block 900 and there are commands by which the CPU can read any location in any of the memories within segmentation block 900. In the case of the "Init FBL" opcode, execution of the command takes approximately eighty cycles of the 200 MHz clock. No data passes through the segmentation block 900 until the initialization is completed.

CRC generator block 909 generates the CRC that covers the PDU (Protocol Data Unit) in the AAL5 format. CRC generator block 909 actually includes two CRC generators, one CRC-32 engine for 64-bit wide data input, and one CRC-32 engine for 32-bit wide data input. The CRC generator block 909 uses the port ID to read a partial CRC from CRC SRAM block 910. CRC SRAM 910 is a 65 x 32 bit dual-port SRAM that stores a partial CRC for each port. The partial CRC is loaded into the CRC-32 engine for 64-bit wide data and the CRC-32 engine for 64-bit wide data calculates the CRC as the data is output to memory manager block 204. All the data up to the AAL5 trailer is passed through this CRC-32 engine for 64-bit wide data input. The partial CRC output by this CRC-32 engine is then loaded into the second CRC engine for 32-bit wide data input. Then the four most significant bytes of the AAL5 trailer are sent through the CRC-32 engine for 32-bit wide data (two of these bytes are the packet length determined by the packet length counter of the input phase state machine block 904). The resulting CRC output by the CRC-32 engine for 32-bit wide data is then the final CRC for the AAL5 trailer. At the end of a packet, the final CRC is multiplexed out by multiplexer 920 to be the last four bytes of the AAL5 trailer.

Output phase state machine block 911 sends a seg_mem_available signal 921 to memory manager block 204 to indicate that a completed 64-byte buffer is available to transfer. Memory manager block 204 replies with the mem_seg_ctrl_pop signal

922 if it is ready to receive the completed 64-byte buffer. When the mem_seg_ctrl_pop signal is received from memory manager block 204, the output phase state machine block 911 reads control queue FIFO 912 for the control information entry for the available 64-byte buffer, reads the appropriate 64-byte buffer from data SRAM 908, and transmits the data from the 64-byte buffer along with the control information to memory manager block 204. Data SRAM 908 is sixty-four bits wide. Individual words of data pass via lines 923, through multiplexer 924, and to pack 128 block 916. Pack 128 block 916 converts two consecutive 64-bit words into one 128-bit word. The 128-bit word is then passed to memory manager block 204 via lines seg_mem_data[127:0]. There are eight time slots of the global 200 MHz clock that clocks the MS-SAR, the current time slot being indicated by a value timeslot[2:0] received via lines 925 from a timeslot generator. The output phase state machine block 911 only outputs information to memory manager block 204 in one predetermined time slot. Memory manager block 204 reads the data from a 64-byte buffer in four consecutive reads of the 128-bit bus.

As set forth above in connection with Figure 53, egress MS-SAR 603 can serially communicate with the segmentation block of an ingress MS-SAR 608 via a serial bus 609 such that the ingress MS-SAR 608 is made to transmit a status cell 611 through the switch fabric 610. The lines pfq_seg_valid, pfq_seg_clk, and pfq_seg_data[2:0] are the serial bus 609 signals. PFQ_SEG interface block 914 receives a twelve-bit status cell in serial fashion via this serial bus. When a status cell is pending in the PFQ_SEG interface block 914, the PFQ_SEG interface block 914 generates a request to output phase state machine block 911. Output phase state machine block 911 then receives the status cell via lines 926 and transmits the status cell using one of the regular port accesses. The transmission of a status cell takes priority over the transmission of other 64-byte chunks to the memory manager block 204.

As seen in Figure 53, status cell information received by egress MS-SAR 606 is communicated to ingress MS-SAR 604 via serial bus 607. In Figure 57, SEG_SCH interface block 915 monitors the lut_seg_type[3:0] lines. If the type is "type 7" when lut_seg_valid signal 919 is asserted, then a status cell is present. The SEG_SCH interface block 915 reads the status cell via lut_seg_data[63:0], serializes the

information, and outputs it as a twelve-bit value via the seg_sch_data[2:0] and seg_sch_valid lines. These lines and the seg_sch_clk clock line comprise the serial bus 607 that extends to the scheduler block of ingress MS-SAR 604.

The CPU can also inject a 64-byte chunk of information. The CPU does this via CPU interface block 913 and a "CPU inject" command. If CPU interface block 913 contains a 64-byte chunk of information that is ready for forwarding to memory manager block 204, then output phase state machine block 911 will output that 64-byte chunk provided there is no status cell pending. The output phase state machine block 911 will only output one CPU-injected chunk once in every sixty-four forwarded 64-byte chunks. When output phase state machine block 911 finishes outputting a 64-byte chunk to memory manager block 204, it returns the pointer to the buffer where the chunk was stored to the free buffer list block 905 via fbl_wr and fbl_addr lines 927.

The SEG_TST multiplexer block 917 is provided for test purposes. It multiplexes one of many test points within the segmentation block 203 onto test terminals (test pins) of the MS-SAR integrated circuit.

MEMORY MANAGER BLOCK IN MORE DETAIL:

Figure 58 is a simplified block diagram of one particular embodiment 1000 of memory manager block 204 of Figure 10. Memory manager block 1000 includes a cell depository manager block 1001, an SRAM adapter block 1002, and a CPU interface block 1003. Cell depository manager block 1001 includes an enqueue control block 1004, a dequeue control block 1005, and a per flow queue interface block 1006. SRAM adapter block 1002 includes a memory controller block 1007 and a cell composition block 1008.

When a 64-byte chunk is ready to be transferred from segmentation block 203 to payload memory 217, segmentation block 203 asserts a SEG_MEM_AVAILABLE signal to enqueue control block 1004. When enqueue control block 1004 detects the SEG_MEM_AVAILABLE asserted, it asserts the MEM_SEG_CNTL_POP signal back to segmentation block 203 in a predetermined time slot. Segmentation block 203 then sends control information including SEG_MEM_FID[19:0], SEG_MEM_TYPE[3:0], SEG_MEM_CLASS[2:0], SEG_MEM_EOP, and

1 SEG_MEM_SOP, to enqueue control block 1004. Segmentation block 203 then
 2 sends sixty-four bytes of data to enqueue control block 1004 as four 128-bit
 3 transfers over SEG_MEM_DATA[127:0]. These four transfers occur in four
 4 predetermined time slots. The data is pushed onto a 128-bit by 64-bit deep data
 5 FIFO in enqueue control block 1004. The control information is formed into a control
 6 word which is supplied to per flow queue block 207 via per flow queue interface
 7 block 1006. Per flow queue interface block 1006 manages the passing of the control
 8 info from the clock domain of memory manager block 1000 to the clock domain of
 9 per flow queue block 207. The control information is passed via lines
 10 MEM_PFQ_PARAM[23:0] along with an enqueue request signal via line
 11 MEM_PFQ_ENQ.

12 Per flow queue block 207 returns a 20-bit buffer ID BID[19:0] via lines
 13 PFQ_MEM_PARAM[23:0]. The buffer ID indicates which 64-byte buffer in payload
 14 memory 217 will be used to store the 64-bytes of data. PFQ_MEM_DEQ indicates
 15 that the BID is being supplied as an enqueue command and not as a dequeue
 16 command. BID[19:0] passes via per flow queue interface block 1006 to enqueue
 17 control block 1004 where it is pushed onto a BID FIFO. The BID FIFO is twenty bits
 18 wide and sixteen bits deep. Each location in the BID FIFO is associated with four
 19 corresponding locations in the data FIFO where the corresponding data is stored.
 20 When a BID is present in the BID FIFO, enqueue control block 1004 asserts an
 21 available signal via an ENQ_CELL_AVAILABLE line to SRAM adapter block 1002.
 22 SRAM adapter block 1002 responds by popping the data FIFO four times and the
 23 BID FIFO once via line ENQ_CELL_DATA_POP such that the BID and associated
 24 data are transferred to memory controller block 1007. Memory controller block 1007
 25 writes to payload memory 217 four times to store the sixty-four bytes of data into
 26 payload memory 217 starting at the location identified by BID[19:0]. In this
 27 embodiment, there is only port for interfacing to external memory 217. Memory
 28 manager block 1007 manages transferring the 64-byte chunk of data in multiple
 29 accesses to external memory 217 via this single port.

30 Although in this embodiment external memory 217 is ZBT (zero bus turnaround)
 31 SRAM, external memory 217 may in other embodiments be another type of memory
 32 such as, for example, DRAM. In some embodiments, bandwidth to external memory

217 is increased by realizing external memory 217 in multiple integrated circuit memory devices, where each is accessed via a different one of a plurality of interface ports controlled by the same memory controller block 1007. Memory controller 1007 in Figure 58 is programmable to interface with pipelined memory. In one such embodiment, the four memory locations where a chunk of data is stored are read in sequence. Although the reading of the first of the four locations requires multiple clock cycles, pipelining is employed such that the other of the four locations are read, one location on each subsequent clock cycle.

After the 64-byte chunk of data is stored in a 64-byte buffer identified by BID[19:0], the 64-byte chunk is retrieved in response to a dequeue command. In a dequeue command, per flow queue block 207 sends a buffer ID BID[19:0] and associated control information to per flow queue interface block 1006 via lines PFQ_MEM_PARAM[23:0]. BID[19:0] indicates where in external memory 217 the data chunk is stored. The PFQ_MEM_DEQ signal is asserted to indicate the BID is supplied as part of a dequeue command. Per flow queue interface block 1006 pushes the BID[19:0] into a BID FIFO in dequeue control block 1005 and the control information into a control FIFO in dequeue control block 1005. Dequeue control block 1005 sends an available signal to memory controller block 1007 via the DEQ_CELL_AVAILABLE line. Memory controller block 1007 responds by popping the BID and control FIFOs in dequeue control block 1005 using the DEQ_CELL_POP signal. Memory controller block 1007 then uses the BID[19:0] to retrieve the identified 64-byte chunk from external memory 217.

23 The retrieved data is passed to cell composition block 1008 via SRAM data line
24 1009 whereas the associated control information is passed to cell composition block
25 1008 via BID/CNTL line 1010. Cell composition block 1008 realigns the control
26 information and the associated data such that the control information and BID are
27 stored at location in a BID/control FIFO that corresponds with four locations in a data
28 FIFO where the data is stored. When corresponding locations in the BID/control
29 FIFO and the data FIFO are loaded, cell composition block 1008 sends an available
30 signal to reassembly block 205 via the MEM_RAS_AVAILABLE line. Reassembly
31 block 205 responds by asserting a RAS_MEM_CONTROL_POP signal to pop the

1 BID and control information from the BID/control FIFO and by asserting a
2 RAS_MEM_DATA_POP signal four times to pop the data from the data FIFO.

3 The CPU can read or write any one of the 64-byte buffers in external memory
4 217 via CPU interface block 1003. CPU interface block 1003 contains sixteen 32-bit
5 data registers and one 32-bit command register. Several bits in the command
6 register indicate an opcode, whereas other bits indicate an address. For example, to
7 write to a 64-byte buffer, the CPU writes the sixteen bytes of data into the sixteen
8 32-bit data registers. The CPU then writes to the command register, the opcode
9 portion of the command register indicating a buffer write operation, the address
10 portion of the command register indicating the BID of the buffer. Once the command
11 register is written, CPU interface block 1003 reads the command register and
12 executes the command. In the case of the write buffer command, the data from the
13 sixteen data registers is written into external memory 217 by memory controller
14 block 1107. CPU interface block 1003 also includes a memory manager
15 configuration register. The CPU can configure memory manager block 1000 in one
16 of numerous ways by writing a configuration word into this configuration register. In
17 one specific embodiment, the configuration register contains: 1) a first multi-bit
18 value which indicates how chip selects to the external memory 217 are to be
19 generated, and 2) a second multi-bit value which indicates an access time of the
20 external memories used to implement payload memory 217.

21

22 REASSEMBLY BLOCK IN MORE DETAIL:

23 Figure 59 is a simplified block diagram of one particular embodiment 1100 of
24 reassembly block 205 of Figure 10. Reassembly block 1100 includes an enqueue
25 state machine block 1101, a free buffer list block 1102, a reassembly table SRAM
26 block 1103, an output port queue block 1104, a control SRAM block 1105, an header
27 SRAM block 1106, a data SRAM block 1107, a dequeue state machine block 1108,
28 a port calendar block 1109, an output FIFO block 1110, an AAL5 CRC checker block
29 1111 containing a CRC table 1112, an L2 CRC checker block 1113 containing a
30 CRC table 1114, a first data pipeline block 1115, a second data pipeline block 1116,
31 a CPU port buffer block 1117, and a CPU interface block 1118.

1 In operation, 64-byte chunks and associated control information are received
2 from memory manager block 204 via 128-bit bus 324 (see Figure 10). Each 64-byte
3 chunk is destined for one of sixty-four possible output ports. The particular output
4 port is identified by an output port ID (PID) received via MEM_RAS_PID[6:0] lines
5 from memory manager block 204. Reassembly block 1100 includes a dual-port
6 SRAM that contains one hundred and twenty-eight 64-byte buffers for storing 64-
7 byte chunks. There is a queue (i.e., a linked list) of pointers to these 64-byte buffers
8 in the reassembly block for each of the active output ports. A 64-byte chunk of data
9 received from memory manager block 204 is stored in a 64-byte buffer and a pointer
10 to the buffer is pushed onto the queue for the particular output port. The output port
11 is designated by MEM_RAS_PID[6:0] for that 64-byte chunk. The flow ID (FID) of
12 the 64-byte chunk is used to lookup header information from header table 327 (see
13 Figure 10) stored in external reassembly memory 218 (see Figure 59). The
14 retrieved header information for the 64-byte chunk is stored in a location in header
15 SRAM 1106 that is also pointed to by the same pointer. In this way, pointers to the
16 64-byte chunks and the associated header information are stored in reassembly
17 block 1100 in per-port queues. Enqueue state machine 1101 handles the
18 enqueueing operation. Dequeue state machine 1108 then pops a selected per-port
19 queue to retrieve a pointer, reads the 64-byte chunk and the header information
20 pointed to by the popped pointer, does processing on the 64-byte chunk in
21 accordance with the application type of the 64-byte chunk, and pushes the
22 processed 64-byte chunk onto output FIFO 1110. The particular per-port queue
23 popped is determined by port calendar 1109. Outgoing SPI-4 interface block 206
24 pops output FIFO 1110 and the processed information is transferred to outgoing
25 SPI-4 interface block 206.

26 More detailed operation of reassembly block 1100 of Figure 59 is now described.
27 Memory manager block 204 contains a first FIFO for storing 64-byte chunks of data.
28 Memory manager block 204 also contains a second FIFO for storing control
29 information associated with the 64-byte chunks of data in the first FIFO. The control
30 information includes an SOP bit, an EOP bit, a twenty-bit flow ID (FID), a seven-bit
31 output port ID (PID), a four-bit application type (TYPE), a CLP bit, an EFCI bit, an
32 OAM bit, and a three-bit CLASS. If such a 64-byte chunk is available in memory

1 manager block 204, then memory manager block 204 sends an available signal to
 2 enqueue state machine 1101 via the MEM_RAS_AVAILABLE line 1119. Enqueue
 3 state machine 1101 requests a pointer to a free 64-byte buffer via request line 1120.
 4 Free buffer list block 1102 returns a pointer to enqueue state machine 1101 via
 5 PTR[6:0] lines 1121. This pointer points to one of 128 buffers in data SRAM 1107,
 6 and also points to one of 128 pairs of locations in header SRAM block 1106, and
 7 also points to one of 128 locations in control SRAM 1105. The 64-byte buffer
 8 pointed to is "free" in that it is not used by any queue (i.e., is free).

9 Once enqueue state machine 1101 has a pointer to a free buffer, it sends a pop
 10 signal to memory manager block 204 to pop the FIFO containing the data via
 11 RAS_MEM_DATA_POP line 1122. A 64-byte chunk then passes via
 12 MEM_RAS_DATA[127:0] lines 1123 and data pipeline 1115 and into data SRAM
 13 block 1107 to be stored at the location indicated by the pointer. Enqueue state
 14 machine block 1101 sends a pop signal to memory manager block 204 to pop the
 15 FIFO containing the associated control information via RAS_MEM_CTRL_POP line
 16 1124. Control information associated with the 64-byte block then passes to
 17 enqueue state machine block 1101 via lines MEM_RAS_SOP, MEM_RAS_EOP,
 18 MEM_RAS_FID[19:0], MEM_RAS_PID[6:0], MEM_RAS_TYPE[3:0],
 19 MEM_RAS_CLP, MEM_RAS_EFCI, MEM_RAS_OAM, and
 20 MEM_RAS_CLASS[2:0]. MEM_RAS_PID[6:0] designates the output port.
 21 MEM_RAS_TYPE[3:0] designates the application type to be performed on the
 22 chunk. MEM_RAS_FID[19:0] designates the flow ID of the chunk.

23 Enqueue state machine block 1101 uses the flow ID to read the associated FID
 24 entry in the header table stored in external reassembly memory 218. Figure 33
 25 illustrates one such FID entry. The FID entry is read by enqueue state machine
 26 block 1101 via HDR_DATA[35:0] lines 1125 and is stored in header SRAM 1106 at a
 27 pair of locations pointed to by the pointer PTR[6:0]. Enqueue state machine block
 28 1101 stores the control information for the chunk in a location in control SRAM 1105
 29 pointed to by the pointer. The same pointer PTR[6:0] on lines 1126 points to the
 30 eight 64-bit locations in data SRAM 1107 that hold the 64-bytes of data, points to the
 31 two 256-bit locations in header SRAM 1106 that stores the FID entry, and points to
 32 the one 32-bit location in control SRAM 1105 that stores the control information.

1 Packet information, regardless of its length, is segmented and stored by the MS-
2 SAR in 64-byte chunks as explained above. By convention, an indication of the
3 length of the packet (in terms of number of valid bytes) is stored in a trailer at the
4 end of the last 64-byte chunk. Accordingly, the last 64-byte chunk of such a packet
5 could contain a little bit of data, followed by padding, and followed by the trailer.
6 Where reassembly block 1100 is to remove the padding, the reassembly block 1100
7 must know how much padding there is. Reassembly block 1100 therefore maintains
8 a "running packet length" count of the total number of bytes (the count includes valid
9 data bytes, as well as any padding and the trailer) received for a packet, from 64-
10 byte chunk to 64-byte chunk, up to and including the last chunk of the packet that
11 has its EOP bit set. The "running packet length" count at the end of the last 64-byte
12 chunk (the last chunk is the chunk with its EOP set) is the number of bytes in the
13 packet. The number of bytes of padding is determined by recovering the packet
14 length value (number of valid data bytes in the packet) from the trailer (the trailer by
15 convention is at the end of the last 64-byte chunk), and subtracting this number of
16 valid data bytes from the "running packet length" count, and then also subtracting
17 the number of bytes of the trailer (the number of bytes of the trailer is a known and
18 fixed number). Accordingly, reassembly block 1100 maintains one such 16-bit
19 "running packet length" byte count for each of the 64 possible output ports. These
20 16-bit running packet length count values are stored in reassembly table SRAM
21 1103 on a per port bases. Before the data of a chunk is written to data SRAM 1107,
22 enqueue state machine block 1101 uses the PID of the 64-byte chunk to read the
23 16-bit running "packet length" partial byte count from reassembly table SRAM 1103.
24 It adds to this count in accordance with how many bytes (including any padding or
25 trailer) there are in the current 64-byte chunk, and then before the next 64-byte
26 chunk is read from memory manager block 204 writes the updated running "packet
27 length" byte count value back into reassembly table 1103 at the location designated
28 by the PID.

29 The per-port queues are maintained using entries in control SRAM 1105. Each
30 32-bit entry in control SRAM 1105 contains the following fields: 1) a one-bit PEND
31 field, 2) a four-bit TYPE field which indicates the application type (type of processing
32 to be done on the associated chunk), 3) a six-bit "cell length" field that indicates the

number of valid data bytes in the associated 64-byte chunk, 4) an L2EN bit which indicates L2 header stripping and CRC checking is enabled, 5) an L2HDR bit which indicates the number of L2 header bytes to be stripped, 6) one BAD bit which indicates an error condition has occurred, 7) one SOP bit, 8) one EOP bit, 9) a seven-bit NEXT RD_PTR field which indicates a pointer to where the next 64-byte chunk in the per-port queue is stored in the reassembly block, 10) a two-bit NEXT HDR_WD field which indicates the number of header bytes for the next 64-byte chunk, and 11) a three-bit MPLS field which indicates what kind of MPLS operation that the reassembly block is to perform. By the use of the NEXT RD_PTR, the entry for one pointer can be made to point to another entry.

Output port queue block 1104 contains sixty-four 24-bit entries, one entry being for each output port. Each entry contains the following fields: 1) one EOP bit, 2) a seven-bit RD_PTR (head pointer) field that indicates where on the reassembly block the 64-byte chunk at the head of the per-port queue is stored, 3) a seven-bit WR_PTR (tail pointer) field that indicates where on the reassembly block the 64-byte chunk at the tail of the per-port queue is stored, 4) one empty status bit which indicates if there are no 64-byte chunks queued for that port, 5) one PEND bit which indicates if a 64-byte chunk is waiting to be queued to the output port, and 6) two HDR_WDS bits which indicate the number of header words for the next 64-byte chunk for that port to be sent to the outgoing SPI-4 interface block.

It is therefore seen that output port queue block 1104 stores, for each per-port queue, a head pointer that points to the head entry in control SRAM 1105 and a tail pointer that points to the tail entry in control SRAM 1105. If the empty bit in an output port queue is set, then the queue is empty. If the head pointer and the tail pointer are the same, then there is only one entry in the queue. If the head pointer points to an entry, the NEXT RD_PTR field of which points to the tail pointer, then there are two entries in the queue. To add a third pointer to a queue, the entry in control SRAM 1105 pointed to by the queue's tail pointer is modified so that its NEXT RD_PTR field points to the entry pointed to by the new third pointer. The tail pointer field for that port in output port queue block 1104 is then changed to be the new third pointer. A 64-byte chunk received from memory manager block 204 is therefore said to be "pushed " onto a per-port queue when the 64-byte chunk is

1 stored in data SRAM 1107 and a pointer to the stored chunk becomes the new tail
2 pointer.

3 Port calendar 1109 identifies an output port as explained above in connection
4 with Figure 24. Dequeue state machine 1108 sends a request to port calendar block
5 1109 via request line 1127 and port calendar 1109 responds by identifying an output
6 port via OUT_PID[6:0] lines 1128. Dequeue state machine block 1108 then pops
7 the per-port queue for that identified output port. It does this by reading the entry for
8 that port in output port queue block 1104, retrieving the head pointer, and then using
9 the head pointer to read: 1) the pointed to 64-byte chunk from data SRAM 1107, 2)
10 the pointed to header information from header SRAM 1106, and 3) the pointed to
11 control information from control SRAM block 1105. The popped head pointer is no
12 longer the head pointer. Rather, the pointer to the next 64-byte block identified by
13 the NEXT RD_PTR field in the control SRAM block 1105 is written into the entry in
14 output port queue block 1104 to be the new head pointer. The old head pointer is
15 communicated to enqueue state machine 1101 via RETURN_PTR[6:0] lines 1129.
16 Enqueue state machine 1101 causes this pointer to be "released" (i.e., be indicated
17 as "free") by sending the pointer to free buffer list block 1102 via PTR[6:0] lines 1130
18 along with a release signal via RELEASE line 1131. The 64-byte chunk of data
19 passes from data SRAM block 1107 via lines 1132, through data pipeline 1116,
20 through multiplexer 1133, and into output FIFO 1110. Output FIFO 1110 is big
21 enough to hold four 64-byte chunks at one time. The header for the 64-byte chunk
22 that is stored in header SRAM 1106 is multiplexed via multiplexer 1133 into the
23 output FIFO so as to be in the appropriate place with respect to the data. Dequeue
24 state machine block 1108 controls multiplexer 1133 via HDR_SEL line 1134.
25 Processing in accordance with the application type indicated by the value TYPE[3:0]
26 read out of control SRAM 1105 is carried out. Which bytes in the 64-byte chunk are
27 padding are indicated by an eight-bit value stored in output FIFO 1110 with each 64-
28 byte chunk. There is one bit in this value RAS_SPIO_BYTE_EN[7:0] for each byte
29 in the 64-byte chunk. Outgoing SPI-4 interface block 206 pops the output FIFO
30 1110 via line SPIO_RAS_POP. Any byte of the received data on lines
31 RAS_SPIO_DATA[63:0] for which the corresponding bit in

1 RAS_SPIO_BYTE_EN[7:0] is not set is not read into outgoing SPI-4 interface block
2 206. The byte of padding is therefore effectively removed.

3 There are two CRC blocks: 1) enqueue CRC generator block 1111 that
4 generates the CRC for the AAL5 format, and 2) dequeue CRC generator block 1113
5 that generates the CRC for the L2 format. The polynomial used is CRC-32. There
6 is a minimum of one clock delay in generating the CRC because the internal data
7 path is sixty-four bits and the CRC polynomial is calculated thirty-two bits at a time.
8 The CRC is generated in two stages, first the upper thirty-two bits and then the lower
9 thirty-two bits. For AAL5 encapsulation, the CRC is checked during the enqueue
10 process of writing the data into data SRAM block 1107. CRC generator block 1111
11 reads a partial CRC using the port ID as the address to CRC table SRAM 1112.
12 CRC table SRAM 1112 stores one partial CRC for each of the sixty-four output
13 ports. Enqueue CRC generator block 1111 loads the partial CRC and starts
14 calculating as the data is coming in from memory manager block 204. When the
15 complete 64-byte chunk is loaded in data SRAM block 1107, enqueue CRC
16 generator block 1111 writes back the partial CRC into CRC table SRAM 1112, again
17 using the port ID as the address. At the end of the last chunk of a packet (as
18 determined by EOP), the CRC is compared to the CRC in the AAL5 trailer. If the
19 compare fails, then the packet is marked as bad by setting a BAD bit in control
20 SRAM 1105 via AAL5_CRC_OK line 1136. When this packet is sent out of
21 reassembly block 1100 via output FIFO 1110, this bad bit is communicated to
22 outgoing SPI-4 interface block 1100 via line 1137, OR gate 1138, the bad bit in
23 output FIFO 1110 for the last 64-byte chunk of the packet, and signal
24 RAS_SPIO_ABORT that is communicated to outgoing SPI-4 interface block 206.

25 For the L2 encapsulation, the CRC is checked during the dequeue process while
26 reading the data out of data SRAM 1107. L2 CRCR checker block 1113 has a CRC
27 table 1114 for storing one partial CRC for each of the sixty-four output ports. This
28 CRC table 1114 is accessed on a per-port basis. The end of the data packet is
29 determined by comparing the packet length calculated and stored in reassembly
30 table SRAM 1103 with the actual packet length (valid bytes) recorded in the AAL5
31 trailer. Once the end of the data is located, the L2 CRC is checked at the calculated
32 packet boundary to determine the integrity of the L2 CRC for that data packet. If L2

1 CRC checker block 1113 determines that the packet has a CRC error, then the error
2 condition is sent to the outgoing SPI interface block 206 via line 1139, OR gate
3 1138, the bad bit in output FIFO 1110, and signal RAS_SPIO_ABORT.

4 CPU interface block 1118 is an interface by which the CPU can cause certain
5 operations to be performed depending on which one of a plurality of opcodes the
6 CPU writes into a 32-bit command register in CPU interface block 1118. Using this
7 32-bit command register and six other general purpose 32-bit registers, the CPU can
8 read from and write to any of the memory locations in any memory in reassembly
9 block 1100. Certain opcodes allow the CPU to read status and statistics information.
10 For example, reassembly block 1100 in some embodiments includes a per port
11 statistics memory (not shown) that includes sixty-four entries, one for each output
12 port. In one embodiment, each entry is 112 bits long and includes: 1) a sixteen-bit
13 TTL_ERRORS field that accumulates a count of the total number of TTL timeout
14 errors detected on that port as various packets are received on that port, 2) a
15 sixteen-bit CRC_ERRORS field that accumulates a count of the number of CRC
16 errors detected on the output port as the various packets are received on that port,
17 3) a 32-bit PKT_RCVD field that accumulates a count of the total number of cells or
18 packets that are received on that output port, and 4) a 48-bit DATA_RCVD field that
19 accumulates a count of the total number of valid bytes received on that output port
20 as the various packets are received on that port. The CPU can read entries from
21 this per port statistics memory via CPU interface block 1118.

22 In the embodiment of Figure 59, the CPU can receive 64-byte data chunks
23 through CPU interface block 1118 using a special output port (output port 65). Data
24 for this special output port is buffered into CPU port buffer block 1117 so that
25 reassembly block 1100 can reassemble the data into the proper configuration before
26 the data is read by the CPU. If output scheduler block 210 sends such a CPU
27 chunk, then reassembly block 1100 sets the CPU FULL status bit and waits for at
28 least sixty-four cycles before clearing the bit. Output scheduler block 210 does not
29 send another CPU chunk until the falling edge of the CPU FULL status bit. Once
30 reassembly block 1100 detects a 64-byte CPU chunk as determined by the PID,
31 reassembly block 1100 writes the 64-bit control word for the CPU chunk into CPU
32 port buffer block 1117. The data of the CPU chunk is then written into CPU port

1 buffer block 1117. After one complete CPU chunk is received, CPU interface block
2 1118 reads the control word from the FIFO. Based on the type field, CPU interface
3 block 1118 determines if another CPU chunk is required before the current CPU
4 chunk can be supplied to the CPU. This would occur for packet traffic types when
5 an EOP must be detected before the second to last cell can be sent. If another CPU
6 chunk is required, then the CPU FULL status bit to the data base block 208 is
7 cleared after a minimum of sixty-four cycles and scheduler block 210 sends another
8 CPU chunk. Once the second CPU chunk is received and the control word has
9 been read to determine whether it is an EOP or not, then the first chunk along with
10 the associated control header is read out of CPU port buffer 1117 and is loaded into
11 the general purpose registers in CPU interface block 1118. Once these registers are
12 loaded, the CPU data bit in the status register is set informing the CPU that a chunk
13 of data is available for reading. The CPU then issues a "GET CPU" command to
14 read from the registers. If there is padding to be stripped based on the type, then
15 the padding is stripped when the data is read out of CPU port buffer 1117 so that the
16 padding is not loaded into the general purpose registers. The cell length field is
17 used to determine how much padding is to be stripped from the end of the
18 associated chunk. No header operation is performed on CPU chunks, so the only
19 padding is at the end of the chunk.

20 Enqueue data pipe block 1115 converts data from 128-bit wide to 64-bit wide for
21 entry into data SRAM 1107. Enqueue data pipe block 1115 also provides the
22 appropriate delay necessary for enqueue state machine block 1101 to obtain the
23 pointer needed to store the chunk into data SRAM 1107. Dequeue data pipe block
24 1116 delays data to the output FIFO block 1110 so that the header can be inserted
25 and the CRC checked for L2 packets before the last word is sent. Header SRAM
26 block 1106 is a 256x64 bit internal dual port SRAM that stores header information on
27 a per-port basis. The memory is organized as 128 sixteen-byte buffers so that two
28 sixteen-byte headers are stored for each output port. Output FIFO block 1110 is a
29 32 x 88 bit FIFO implemented using dual port memory. The output FIFO is large
30 enough to store three maximum size chunks. The output FIFO has two full
31 indications (Almost Full and Full) and two empty indications (Almost Empty and
32 Empty). Almost Full is asserted when the FIFO has ten locations left. Full is

1 asserted when the FIFO has nine or fewer locations left. By indicating that the FIFO
2 is almost full when it has only ten locations left, the dequeue state machine block
3 1108 can then send one more complete chunk. When the FIFO is one word beyond
4 Almost Full, then the Full signal is asserted thereby indicating that once the current
5 chunk is completed that no more chunks are to be transferred until Full goes
6 inactive. Almost Empty is asserted when one word is left in the output FIFO. Empty
7 is asserted when the output FIFO is empty. Outgoing SPI interface block 206 ORs
8 the two empty signals together to determine when to pop the output FIFO. Outgoing
9 SPI interface block 206 also uses the two empty signals to determine when only one
10 word is left in the FIFO. This is necessary because the minimum POP size is two
11 clock cycles, but when only one word is left the FIFO only has valid data for the first
12 read of the FIFO. If Empty is active but Almost Empty is inactive, then only one
13 word is in the FIFO. Dequeue state machine block 1108 causes data to be loaded
14 into output FIFO block 1110 as long as output port calendar block 1109 indicates
15 chunks are available and output FIFO block 1110 is not full. If output FIFO block
16 1110 goes full, then the current chunk being transferred is completed before
17 reassembly block 1100 stops sending data to output FIFO block 1110. Outgoing
18 SPI interface block 206 pops data from the output FIFO block 1110 as long as FIFO
19 block 1110 is not empty and outgoing SPI interface block 206 is not inserting control
20 words into the data stream. FIFO block 1110 allows outgoing SPI interface block
21 206 to control the flow of data from reassembly block 1100 so it has time to add the
22 control words into the output data stream (coming out of outgoing SPI interface block
23 206) between bursts.

24 External reassembly memory 218 (see Figure 10) is 4Mx36 bits of external
25 SRAM that stores the header control information on a per-FID basis. External
26 reassembly memory 218 includes the appropriately programmed new header for the
27 particular flow and/or the necessary MPLS manipulation control information for that
28 flow. Reassembly memory 218 supports 1M flows with sixteen bytes of header and
29 two bytes of control information per flow. In one embodiment, the MS-SAR operates
30 without any external reassembly memory 218. In such an embodiment, only the
31 Azanda switch header is added in the ingress mode and no MPLS translation occurs
32 in the egress mode. The Azanda header is obtained from side band signals coming

1 from memory manager block 204 and any L2 stripping information is obtained from
2 the control register. For AAL5 type traffic, lookup block 202 (see Figure 10) adds the
3 appropriate ATM header and reassembly block 1100 passes the ATM header on
4 without translation.

5 Dequeue state machine block 1108 performs any MPLS manipulation required
6 pursuant to requirements of the application type. The FID entry in the header table
7 in external memory 218 contains the MPLS label operation instruction and new
8 labels. Three bits in the first control word contain encoded instructions for what
9 MPLS operation to perform. In the embodiment of Figure 59, the operation is either:
10 1) replace the MPLS tag, 2) push the MPLS tag once or twice, or 3) pop the MPLS
11 tag once or twice. After each MPLS operation, the original MPLS TTL value is
12 decremented. If the TTL value is one before being decremented, then the packet is
13 dropped. This is done by not pushing the packet onto output FIFO 1110 and by
14 dropping subsequent chunks from the flow until the EOP is reached. If two labels
15 are being pushed, then label #1 is pushed on first and label #2 is pushed on second.
16 If one label is pushed, then label #1 is used. If two labels are popped, then the third
17 label is the one remaining on the top. If one label is popped, then the second label
18 remains on top.

19 Although the simplified diagram of Figure 59 illustrates many of the important
20 interconnections between the various sub-blocks within the specific embodiment of
21 reassembly block 1100, there are multiple other interconnections not shown in
22 Figure 59 that are used to implement various details of the processing carried out by
23 reassembly block 1100 in accordance with the multiple application types. These
24 many interconnections are omitted from Figure 59 in order to clarify the above
25 explanation of the general operation of reassembly block 1100. Reassembly block
26 1100 is implemented, in accordance with one embodiment, by describing the
27 functionality to be performed in accordance with each of the application types using
28 a hardware description language such as Verilog. Once the functional of the
29 reassembly block has been described in Verilog, circuitry is synthesized from that
30 description using commonly available software tools.

31 Figures 60A-60J illustrate functionality of the reassembly block in certain
32 application types. Because reassembly block operation in several application types

FOOTER 0037621-1030

1 is similar, functional description of the reassembly block 1100 is simplified by
2 describing reassembly block operation in eight different reassembly types.

3 Figure 60A illustrates reassembly type 1 processing. In reassembly type 1, 64-
4 byte chunks in payload memory 217 are mapped into switch cells by adding a switch
5 header. Reassembly type 1 handles reassembly processing for application types 0-
6 3 and 7. Figure 60A shows a one-to-one mapping of ATM cells stored in 64-byte
7 memory buffers into switch cells. The switch header added is eight or sixteen bytes
8 long with the Azanda header being the last four bytes. Reassembly block 1100 in
9 reassembly type 1 informs the SPI interface of the number of valid bytes in each
10 eight-byte word and, for packet traffic, indicates which chunk is the SOP and which
11 chunk is the EOP. For AAL5 cells, the cells are stored in payload memory as 48-
12 byte AAL5 cells with an AAL5 trailer at the end and padding to fill each 64-byte
13 buffer up to sixty-four bytes. For packet traffic, the 64-byte chunks are stored in
14 payload memory as 64-byte chunks with an AAL5-like trailer added to pass on the
15 packet length information. Figure 60B shows AAL5 encapsulated cells. Figure 60C
16 shows pseudo-AAL5 encapsulated packets. For all ingress traffic manager types,
17 the CLASS, OAM, EFCI and CLP bits from memory manager block 204 are inserted
18 into the Azanda four-byte header. For application types 0 and 1, the CLASS, OAM,
19 EFCI and CLP bits are whatever was in the ATM header. For application types 2
20 and 3, the CLASS, OAM, EFCI and CLP are always zero coming from memory
21 manager block 204.

22 Figure 60D illustrates reassembly type 2 processing. In reassembly type 2, 52-
23 byte ATM cells stored as 64-byte chunks are mapped into packets and a packet
24 header is added before the packet is sent on to the outgoing SPI-4 interface block.
25 Only one ATM cell is mapped into each packet. Eight of the twelve bytes of padding
26 at the end of each 64-byte chunk is removed before the packet is sent out. This
27 allows a 64-byte packet to be sent out that includes eight bytes of header. If the
28 switch packet header is sixteen bytes, then the packet is seventy-two bytes long.
29 The CLASS, OAM, EFCI and CLP bits are passed in the Azanda header portion of
30 the switch packet header.

31 Figure 60E illustrates reassembly type 3 processing. In reassembly type 3, 48-
32 byte ATM cells stored in payload memory are mapped into packets by stripping the

1 sixteen bytes of padding from each 64-byte chunk and then optionally stripping the
2 L2 packet header and optionally adding a switch packet header before sending the
3 resulting switch packet on to the outgoing SPI-4 interface block. Reassembly block
4 1100 checks the L2 CRC before stripping the L2 header to ensure packet integrity
5 before sending the packet on to the switch fabric. Reassembly block 1100 also
6 checks against the maximum MTU and the length and CRC in the AAL5 trailer and
7 removes the AAL5 encapsulation. The CLASS, OAM, EFCI and CLP bits are
8 passed in the Azanda header portion of the switch packet header. The L2 CRC
9 being sent out is now incorrect due to the removal of the L2 header and must be
10 replaced by the egress MS-SAR when it adds a new L2 packet header. Lookup
11 block 202 replaces the L2 header and segmentation block 203 calculates the new
12 CRC. For the L2 stripping case, the special header required by lookup block 202 in
13 the egress MS-SAR is added by the switch fabric.

14 Figure 60F illustrates reassembly type 4 processing. In reassembly type 4,
15 reassembly block 1100 maps AAL5-like switch cells stored in payload memory into
16 packets, checks against the maximum MTU, removes the AAL5 encapsulation, and
17 may add an eight or sixteen-byte switch packet header. The CLASS, OAM, EFCI
18 and CLP bits are passed in the Azanda header portion of the switch packet header.
19 Reassembly type 4 can also be carried out with a lookup operation.

20 Figure 60G illustrates reassembly type 5 processing. In reassembly type 5,
21 reassembly block 1100 translates the ATM header by replacing the 28-bit VPI/VCi
22 and then replacing the CLP bit with the logical OR of the CLP coming from the
23 memory manager block 204 and the CLP already in the ATM header before the cell
24 is sent out. The PTI field is left unchanged. The extra twelve bytes of padding from
25 the end of each ATM cell is stored as 64-byte chunks in payload memory. This
26 padding is stripped by the outgoing SPI-4 interface block using byte enables. The
27 remaining information is sent out as a 52-byte ATM cell.

28 Figure 60H illustrates reassembly type 6 processing. In reassembly type 6,
29 reassembly block 1100 maps the 48-byte AAL5 encapsulated ATM cells stored as
30 64-byte chunks in payload memory into packets by stripping the sixteen bytes of
31 padding from each chunk and then performing an MPLS operation based on the
32 contents of the external reassembly memory 218. The resulting packet is then sent

1 on to the outgoing SPI-4 interface block. Part of the MPLS operation is
2 decrementing the MPLS TTL. If the MPLS TTL is one before being decremented,
3 then the packet is dropped. Reassembly block 1100 also checks against the
4 maximum MTU and the length and CRC in the AAL5 trailer before removing the
5 AAL5 encapsulation.

6 Figure 60I illustrates reassembly type 7 processing. In reassembly type 7,
7 reassembly block 1100 adds the four-byte ATM header along with four bytes of
8 padding, so the data is eight-byte word-aligned to the 48-byte AAL5 encapsulated
9 ATM cells stored in payload memory as 64-byte chunks. Reassembly block 1100
10 strips the sixteen bytes of padding, replaces the PTI and CLP bits in the ATM header
11 using the incoming EOP, OAM, CLP and EFCI bits from memory manager block
12 204, and then transmits the chunks as AAL5 cells. The four bytes of alignment
13 padding are stripped using the byte enables when transmitting to the outgoing SPI-4
14 interface block. The EOP AAL5 cell is marked using the PTI field in the ATM
15 header. The packet length is checked against the maximum MTU. The length and
16 CRC in the AAL5 trailer are checked.

17 Figure 60J illustrates reassembly type 8 processing. In reassembly type 8,
18 reassembly block 1100 maps the AAL5 like switch cells stored in payload memory
19 into packets and performs an MPLS operation based on the contents of external
20 reassembly memory 218. Part of the MPLS operation is decrementing the MPLS
21 TTL. If the MPLS TTL is one before being decremented, then the packet is dropped.
22 Reassembly block 1100 also checks the packet length against the maximum MTU.
23 The length and CRC in the AAL5 trailer are checked before removing the AAL5
24 encapsulation.

25 OUTGOING SPI-4 INTERFACE BLOCK IN MORE DETAIL:

26 Figure 60 is a simplified block diagram of one particular embodiment 1200 of
27 outgoing SPI-4 interface block 206 of Figure 10. Port calendar state machine 1201
28 utilizes a port calendar memory 1202 to identify an output port that can receive data.
29 Reassembly block 205 supplies data and control information at a 200 MHz rate via
30 RX state machine 1203 and per port FIFO 1204 to data FIFO 1205. TX state
31 machine 1206 causes the data and control information to be output from data FIFO
32

1 1205 at a 400 MHz double data rate. The data flows through 64-to-16 unpack
2 engine block 1207 which unpacks the 64-bit data words to form 16-bit data words.
3 Data parity generation block 1208 calculates a four-bit diagonal interleaved parity
4 (DIP-4) on the data. TX state machine block 1206 combines the DIP-4 bits with the
5 control information read out of data FIFO 1205 to form 16-bit control words.
6 Multiplexer 1209 multiplexes the control words and data words in accordance with
7 the SPI-4 specification and the resultant 16-bit flow of information passes out of the
8 MS-SAR and onto an external 16-bit, 400 MHz DDR bus. Signal names in Figure 60
9 are descriptive of their information content and function. Operation of port calendar
10 1201 is similar to the operation of the port calendar described in connection with
11 Figure 24.

12 13 CPU INTERFACE BLOCK IN MORE DETAIL:

14 Figure 61 is a simplified diagram of one particular embodiment 1300 of CPU
15 interface block 211 of Figure 10. CPU interface block 1300 provides a general
16 purpose CPU interface port for the MS-SAR integrated circuit. The CPU interface
17 port has a 32-bit data bus. The various registers within the CPU interface blocks of
18 the various other blocks of the MS-SAR are accessible by the CPU through CPU
19 interface block 1300. CPU interface block 1300 receives a ten-bit address from the
20 CPU and decodes it. Three bits of the ten-bit address identify one of the blocks
21 within the MS-SAR whereas the other seven bits identify a register within that block.
22 CPU interface block 1300 generates block select signals that are supplied to each of
23 the other blocks of the MS-SAR.

24 There are also registers within the CPU interface block 1300 that can be read
25 from and/or written to by the CPU. These registers include a read-only version
26 register, a mode register, a soft reset register, and an input/output disable register.
27 The mode register includes: 1) a six bit value which indicates the port number of the
28 MS-SAR when the MS-SAR is in the ingress mode, 2) a one-bit value which
29 indicates whether the MS-SAR is in the ingress mode or in the egress mode, 3) a
30 one-bit value which indicates whether or not the MS-SAR is operating with an
31 external reassembly memory 218, and 4) a four-bit value which selects the test
32 multiplexer of one internal block, the output of which is coupled to the

1 MX1_TEST_OUT[31:0] output test port. Each of these test multiplexers has a thirty-
2 two bit output bus. Test multiplexer 917 of Figure 57 is an example of one such test
3 multiplexer. CPU interface block 1300 controllably multiplexes the 32-bit output bus
4 of one of these test multiplexers onto 32-bit output test bus MX1_TEST_OUT[31:0].

5 The soft reset register allows the CPU to soft reset any individual block within the
6 MS-SAR. The input/output disable register contains bits which indicate for each of
7 the blocks of the MS-SAR whether the input of the block is enabled or disabled, and
8 whether the output of the block is enabled or disabled.

9 CPU interface block 1300 also generates a global synchronization signal
10 GSYNC that is supplied to all blocks. GSYNC is a pulse signal that is asserted once
11 every 520 cycles of the global 200 MHz clock. GSYNC is used to synchronize
12 communication between the various blocks of the MS-SAR. Each block uses
13 GSYNC to generate the three-bit time slot count value (a three-bit value that has
14 eight as its maximum count value) by counting the GSYNC signal.

15
16 Although the present invention is described in connection with certain specific
17 embodiments for instructional purposes, the present invention is not limited thereto.
18 Although the bus protocol by which the incoming bus interface block communicates
19 off-chip and the protocol by which the outgoing bus interface communicates off-chip
20 are the same (SPI-4) in the above-described embodiments, they need not be the
21 same. Only one or neither of the interface blocks need communicate off-chip in
22 accordance with the SPI-4 protocol. The SPI-4 bus is mentioned just as an
23 example. More than one incoming bus interface block can be provided so that the
24 MS-SAR can interface to either of two different types of buses. Although the lookup
25 block is disposed in the data path between the incoming bus interface block and the
26 segmentation block, this need not be the case in all embodiments. The lookup
27 block may, for example, analyze data passing from the incoming bus interface to the
28 segmentation block without acting as a buffer in the data path between the incoming
29 bus interface block and the segmentation block. In some embodiments, the
30 segmentation block does buffer an entire packet before segmenting it, rather than
31 receiving the packet in smaller pieces and then processing those smaller pieces one
32 by one. Similarly, the reassembly block in some embodiments does reassemble an

[illegible]